SoK: Trust Signals in AI-Mediated Commerce
How Generative Engines Select and Rank Products

Kei-Leo Brousmiche[1]
[1]Keyban kei@keyban.io

Preprint — March 2026. This paper has not yet been peer-reviewed.

Abstract. The rise of AI shopping agents and generative engine optimization (GEO) is fundamentally restructuring product discovery. As generative engines replace ranked link lists with synthesized recommendations, the signals that determine product visibility are shifting, yet no systematization exists at the intersection of trust signals, generative engine discovery, and agentic commerce. This paper presents a seven-category taxonomy of trust signals (spanning stylistic optimization, platform badges, social proof, structured data, evidence/citations, traditional certifications, and cryptographic attestations) evaluated across seven dimensions: fabrication cost, machine-readability, standardization, GEO evidence, cross-platform portability, forgery cost, and legal deterrence. Synthesizing 38 references across GEO research, platform economics, and decentralized identity, the analysis reveals a structural *trust signal gap*: the signals that generative engines currently reward are precisely those most vulnerable to fabrication, while the signals most resistant to manipulation remain untested in AI discovery contexts. Cryptographic attestations—particularly W3C Verifiable Credentials—emerge as the most architecturally promising candidate to bridge this gap, combining machine-readability with cryptographic non-repudiation (an approach reinforced by the EU Digital Product Passport regulation, which will mandate machine-readable, verifiable product attestations from 2027). Significant challenges in empirical validation, scalability, and interoperability remain; a six-challenge research agenda is proposed.

1 Introduction

The transition from Search Engine Optimization (SEO) to Generative Engine Optimization (GEO) [1] is reshaping how products are discovered, evaluated, and purchased online. Rather than presenting ranked lists of links, generative engines synthesize unified responses that draw on multiple sources, applying fundamentally different selection logic than traditional search [36, 20]. AI shopping agents (deployed by OpenAI [2], Google [3], and Perplexity [4]) now autonomously evaluate products on behalf of consumers, with 73% of consumers willing to delegate product discovery to AI assistants [29]. Yet 85% of consumers demand explicit control over data collected by AI agents [29, 28]. Two agentic commerce protocols (ACP [13] and UCP [14]) are establishing the infrastructure for machine-to-machine transactions.

This shift creates a trust crisis. AI-generated product description edits yield market share gains in 33% of experiments [7], LLM-generated fake reviews are indistinguishable to human judges at 50.8% accuracy [6], and LLMs exhibit a "veracity bias" that compounds their vulnerability to manipulation [5]. As the cost of producing convincing fake product content approaches zero, the marketplace risks the "lemons problem" described by Akerlof [24]: without reliable quality signals, markets deteriorate.

Despite the growing body of work on GEO strategies, AI shopping agent behavior, and trust in AI recommendations, no existing systematization covers the intersection of *trust signals*, *generative engine discovery*, and *agentic commerce*. Prior work addresses these domains separately: GEO optimization strategies [1, 11, 12], AI agent manipulation [7], consumer trust in AI [28], structured data adoption [30], and decentralized identity systems [33, 35]. This paper bridges these research streams.

This systematization makes five contributions:

1. A **seven-category taxonomy** of trust signals in AI-mediated commerce, spanning stylistic optimization, platform badges, social proof, structured data, evidence/citations, traditional certifications, and cryptographic attestations (Section 3).
2. A **seven-dimension evaluation framework** assessing each category on fabrication cost, machine-readability, standardization, GEO evidence, cross-platform portability, forgery cost, and legal deterrence (Section 5).
3. A **synthesis of 38 references** spanning GEO research, platform economics, structured data, and decentralized identity, identifying cross-cutting empirical patterns (Section 4).
4. The identification of a **trust signal gap**: the signals that generative engines currently reward are precisely those most vulnerable to fabrication, while the signals most resistant to manipulation remain untested in GEO contexts (Section 5).
5. A **six-challenge research agenda** covering empirical validation, platform-side verification, scalability, interoperability, market concentration bias, and regulatory drivers (Section 7).

The remainder of this paper is organized as follows. Sec-

tion 2 establishes the GEO landscape, agentic commerce ecosystem, and manipulation threat. Section 3 presents the seven-category taxonomy. Section 4 synthesizes cross-cutting empirical patterns. Section 5 provides a comparative analysis with a central evaluation table. Section 6 examines how cryptographic attestations address the identified gap. Section 7 outlines the research agenda. Section 8 concludes.

2 Background

This section establishes the technical and economic context for the trust signal landscape analyzed in subsequent sections.

2.1 Generative Engine Optimization

Generative Engine Optimization (GEO) was formally defined by Aggarwal et al. [1] as the practice of optimizing content to improve its visibility in AI-generated responses. Unlike traditional SEO, which targets ranked lists of links, GEO targets the synthesized outputs of generative engines that combine retrieval-augmented generation (RAG) with LLM inference to produce unified responses drawing on multiple sources. The foundational GEO study established a clear hierarchy: strategies providing verifiable evidence (citations, statistics, authoritative quotations) significantly outperform purely stylistic strategies such as keyword stuffing or authoritative tone.

Subsequent work has extended these findings to e-commerce. The E-GEO benchmark [11] evaluated GEO strategies using 7,000+ consumer queries against approximately 52,000 products, demonstrating a "universally effective" GEO strategy based on substantive information enrichment. The CORE framework [12] confirmed that product content characteristics strongly influence generative engine recommendations, achieving high success rates in promoting specific products within LLM search outputs.

2.2 AI Shopping Agents and Agentic Protocols

The emergence of AI shopping agents represents a structural shift in product discovery. ChatGPT Shopping uses a model trained via reinforcement learning for shopping tasks, producing organic, unsponsored product results ranked on relevance, price, ratings, and availability [2]. Google AI Mode applies different ranking logic than traditional search: 47% of AI Overview citations come from pages ranking below position #5 [18], emphasizing semantic completeness, E-E-A-T signals[1], and structured data markup over domain authority. Perplexity Shopping combines intent matching, schema completeness, price/stock freshness, and review trust to select products [4].

[1]Experience, Expertise, Authoritativeness, and Trustworthiness—Google's quality rater guidelines for evaluating content credibility.

Consumer survey data quantifies the scale of this transition. A cross-market study of 3,700 respondents [29] finds that 73% of consumers would delegate product *discovery* to an AI assistant and 69% would delegate *evaluation*.

Two major protocols now define how AI agents interact with merchants. The Agentic Commerce Protocol (ACP), co-developed by OpenAI and Stripe [13], enables structured conversations between buyer agents and businesses, with merchants pushing product feeds directly to OpenAI for discovery via ChatGPT. The Universal Commerce Protocol (UCP), co-developed by Google with Shopify, Target, Walmart, and others [14], provides Discovery, Checkout, and Orders capabilities with interoperability across Agent2Agent (A2A) and Model Context Protocol (MCP), with product data flowing through Google Merchant Center.

Across platforms, several signals are consistently valued: structured data completeness [19], product identifiers (GTIN, MPN) for disambiguation, real-time accuracy of stock and pricing, and institutional credibility signals. Research on 55,936 queries confirms that 37% of domains cited by LLM search engines are unique to LLM-based systems [20], establishing that AI discovery creates visibility pathways independent of traditional SEO.

2.3 The Manipulation Crisis

The susceptibility of AI shopping agents to manipulation constitutes a foundational problem for the trust signal landscape. Allouah et al. [7] demonstrate that AI-generated description edits yield market share gains in 33% of experiments (+3.66 to +14.89 percentage points), with strategic keyword front-loading exceeding 80 percentage points in one category. Platform endorsement badges ("Overall Pick") increase selection rates from a 10% baseline to approximately 20–43% depending on the model, while "Sponsored" badges reduce selection to 7.9–8.9%.

The threat extends to product reviews. LLM-generated fake reviews are indistinguishable to both human judges (50.8% detection accuracy, effectively chance level) and general-purpose LLMs [6]. While supervised detection using fine-tuned models such as DeBERTa-v3 achieves 96–98% accuracy [5], this requires specialized training that is not available in production AI shopping agents. LLMs exhibit a "veracity bias" (a systematic tendency to accept content as truthful), compounding the vulnerability.

This manipulation landscape creates what Akerlof [24] characterized as a market for lemons: as the cost of producing convincing fake product content approaches zero, quality signals that cannot be easily fabricated become increasingly valuable for differentiation. The taxonomy presented in Section 3 evaluates each signal category's resistance to this threat.

3 Taxonomy of Trust Signals

This section presents a seven-category taxonomy of trust signals relevant to AI-mediated commerce. Each category is evaluated along seven dimensions analyzed in Section 5: fabrication cost, machine-readability, standardization, GEO evidence, cross-platform portability, forgery cost (the technical difficulty of producing a fake signal), and legal deterrence (the juridical risk of doing so).

3.1 (a) Stylistic Optimization

Stylistic optimization encompasses modifications to content tone, structure, and presentation aimed at improving generative engine visibility. [1] tested nine such strategies, finding that purely stylistic approaches (authoritative tone, keyword stuffing, fluency optimization) yield minimal to zero improvement in visibility. The E-GEO benchmark [11] confirms this pattern in e-commerce: while a "universally effective" optimization strategy exists for product content, it relies on substantive information enrichment rather than surface-level rewording.

However, content *structure* matters. The CORE framework [12] achieves 80–91% success rates in promoting products within LLM outputs by optimizing content characteristics, and a study of 129,000 domains [37] finds that optimal section length (120–180 words) and content length (2,900+ words) correlate with higher citation rates. Notably, the same study reveals that high keyword semantic matching *reduces* citations (0.84–1.00 similarity scores yield only 2.7 citations vs. 6.4 for low-match content), indicating that generative engines penalize over-optimization, a natural anti-manipulation mechanism absent in traditional SEO.

Stylistic signals score very low on fabrication cost (any merchant can rewrite content), machine-readability (unstructured text), standardization (no formal standard), forgery cost (trivially replicable), and legal deterrence (none). Their GEO evidence is mixed: structural optimization helps, but tone and keyword strategies are ineffective or counterproductive.

3.2 (b) Platform-Managed Badges

Platform-managed badges are algorithmically assigned endorsements controlled by marketplace operators. In agentic contexts, AI agents respond strongly to badges: "Overall Pick" labels increase agent selection rates from a 10% baseline to approximately 20–43% [7]. A controlled experiment across GPT-4.1, Claude Sonnet 4, and Gemini 2.5 Flash confirms that all three frontier models shift product choices in the same direction when endorsement badges are present, though effect magnitudes vary sharply by model [8]. Critically, both studies measure agent behavior *within* a single storefront; no evidence exists that platform badges influence product selection in cross-platform GEO contexts (e.g., ChatGPT Shopping or Perplexity citing badge-holding products).

Despite their effectiveness, platform badges have critical limitations for AI-mediated commerce: they are not exposed as structured data (requiring HTML parsing), follow no cross-platform standard, and offer zero portability (an endorsement on Amazon provides no signal to ChatGPT Shopping or Perplexity). Their forgery cost is high (algorithmically assigned, not purchasable), but this comes at the cost of complete vendor lock-in.

3.3 (c) Social Proof and Reviews

Social proof encompasses consumer reviews, ratings, and user-generated content that signal product quality through collective experience. This category faces a fundamental integrity crisis in AI-mediated contexts. LLM-generated fake reviews are indistinguishable to human judges, who achieve only 50.8% detection accuracy (chance level), and general-purpose LLMs perform no better [6]. While specialized detection models achieve 96–98% accuracy [5], production AI shopping agents lack this capability.

The role of social proof in AI search diverges sharply from traditional commerce. A comparative study across five platforms [36] finds that social content (Reddit, forums, user reviews) accounts for 0% of ChatGPT citations in most product categories, compared to 10–23% in traditional Google Search. AI engines systematically prefer earned media (third-party editorial content) over user-generated social proof. However, *aggregated* review signals on third-party platforms retain influence: domains present on review platforms such as Trustpilot, G2, and Capterra receive approximately 3× more ChatGPT citations than those absent from such platforms [37].

This creates a paradox: individual reviews are increasingly unreliable and ignored by AI search, but aggregated review presence on trusted third-party platforms serves as a credibility signal. Social proof scores high on fabrication cost (building legitimate review volume requires real customers and sustained product quality over time), medium on machine-readability (AggregateRating schema exists but individual review authenticity is unverifiable), partial on standardization (Schema.org Review types), and very low on forgery cost (the cost of generating fake reviews approaches zero) with negligible legal deterrence.

3.4 (d) Structured Data and Knowledge Graphs

Structured data (primarily Schema.org markup in JSON-LD format) provides machine-readable product information that AI systems can directly parse and reason over. Web-scale analysis of 16.9 million sites [30] reveals that JSON-LD adoption has reached 41% (+7% year-over-year), making it the fastest-growing structured data format. However, Product-type markup remains at only 0.77% of pages despite 3.1 million implementations, indicating a significant adoption gap in e-commerce.

The impact on visibility is substantial. Rich results generated from structured data achieve a 58% click-through rate compared to 41% for non-rich results across 4.5 million queries [31]. More fundamentally, knowledge graph representations improve LLM accuracy by 3.4× compared to unstructured data (from 16% to 54% on enterprise question-answering tasks), with "schema-intensive" questions achieving 0% accuracy without structured representations [32]. Industry guidance confirms this direction: AI shopping agents "prioritize structured, machine-readable data: product information, pricing, availability, and clear policies" [29].

Structured data scores very high on machine-readability (by definition), high on standardization (Schema.org is the de facto universal vocabulary), and high on portability (web-native, accessible to any crawler). However, structured data scores very low on forgery cost: markup is self-declared by the website with no third-party verification. A merchant can annotate any product with arbitrary structured claims (including false certifications, inflated ratings, or fabricated availability) with no cryptographic proof of accuracy. This vulnerability is the central motivation for the verifiable credential approach analyzed in Section 6.

3.5 (e) Evidence and Citations

Evidence-based signals (embedded citations, statistical data points, expert quotes, and references to authoritative sources) represent the strongest driver of generative engine visibility identified in the literature. [1] found that adding citations improves visibility by 30% overall and up to 115% for lower-ranked sites, while adding statistics yields +35% and authoritative quotations +40%. A large-scale analysis of 129,000 domains [37] corroborates this hierarchy: content with 19+ data points receives 93% more ChatGPT citations than data-sparse content, and pages with expert quotes receive 71% more citations.

However, evidence-based signals are susceptible to source framing effects. An experimental study of 192,000 assessments across four LLMs [22] demonstrates that source attribution systematically alters LLM evaluation of identical content: models achieve over 90% inter-model agreement when no source is attributed, but alignment breaks down when text is attributed to authors of different nationalities. This indicates that LLMs do not evaluate evidence purely on merit; perceived source authority modulates the signal's effectiveness. An analysis of 615 health sources cited by ChatGPT [23] confirms this pattern: over 75% of citations come from established institutional sources, suggesting that evidence carries more weight when paired with institutional authority.

Evidence signals score medium on fabrication cost (producing genuine data and expert commentary requires real investment, but citations can be fabricated), medium on machine-readability (citations are embedded in text, not structured), and low on forgery cost (fake statistics and fabricated expert quotes are possible), though fabricating attributed claims from named institutions carries moderate legal deterrence. They have no formal standard and medium portability, as embedded evidence travels with the content across platforms.

3.6 (f) Traditional Certifications

Traditional certifications (ISO standards, organic labels, fair trade marks, safety certifications) represent established trust mechanisms with decades of institutional backing. The critical question for this paper is whether certifications influence AI agent product selection, and here, to the best of our knowledge, **no published study has directly tested this**. The closest evidence comes from [26], which demonstrates that textual trust labels (social proof, authority cues) systematically shift LLM product recommendations, with effects persisting across model scales and even chain-of-thought reasoning. Certifications are functionally analogous to the authority labels tested, suggesting they *could* influence LLM rankings, but this remains an untested hypothesis. This absence of GEO evidence is reflected in the "None" rating in Table 1.

Traditional certifications also face a fundamental machine-readability barrier: they are typically represented as visual badges (images) with no structured, machine-parseable encoding. An AI agent cannot programmatically verify whether a product holds ISO 14001 certification from a webpage image alone. This scores certifications low on machine-readability, low on portability (platform-specific display), but very low on forgery cost (any merchant can textually claim a certification with no machine-verifiable proof). Their primary deterrent is legal: falsely claiming a protected certification mark carries trademark infringement and consumer fraud liability, scoring high on legal deterrence. This asymmetry (technically trivial to fake, legally risky to fake) is precisely the gap that cryptographic attestations address.

3.7 (g) Cryptographic Attestations

This category encompasses trust signals whose credibility derives from cryptographic proof rather than platform authority or self-declaration. The unifying property is that any third party can mathematically verify the authenticity, integrity, and provenance of a claim without trusting the channel through which it was received. Several approaches exist: on-chain data anchoring (publishing signed product claims or their hashes to a blockchain, providing tamper-evident timestamping), digitally signed metadata (e.g., JSON-LD payloads signed with a manufacturer's key), and—most prominently—W3C Verifiable Credentials (VCs).

VCs, published as a W3C Recommendation in May 2025 [10], represent the most mature and standardized form of cryptographic attestation. A VC binds claims about a subject to a cryptographic proof from an identified issuer, whose identity is anchored by a Decentralized Identifier

(DID) [9]—a globally unique URI resolved through a decentralized registry (typically a blockchain or distributed ledger) that returns the issuer's public keys and verification methods. This enables any verifier to confirm authenticity and integrity without contacting the issuer. VCs support multiple serialization formats (JSON-LD, JWT, CBOR) and privacy-preserving features such as selective disclosure via SD-JWT. A comprehensive survey [33] documents multiple open-source implementations (DIDKit, IOTA Identity, Hyperledger Aries, Microsoft Entra, Veramo), confirming that the tooling ecosystem is maturing.

Early empirical assessment [34] shows that AI agents can technically handle DIDs and VCs (authenticate, issue, verify credentials), though with widely varying completion rates and significant cost and latency barriers for real-time commerce. A critical finding is that LLMs sometimes alter VC content during processing, breaking cryptographic integrity, indicating that verification must remain in a deterministic layer rather than being delegated to the LLM. The broader deployment landscape remains nascent, with high attrition among early decentralized identity systems and government-backed initiatives dominating adoption [35]. However, regulatory momentum is building: the EU Digital Product Passport (DPP) regulation will mandate machine-readable, verifiable product attestations from 2027 for priority product categories, and the UN Transparency Protocol (UNTP), adopted as UN/CEFACT Recommendation No. 49, prescribes W3C VCs as the credential format for Digital Product Passports in global supply chains [15, 16].

Cryptographic attestations score very high on machine-readability (JSON-LD by design for VCs, structured by construction for on-chain anchoring), standardization (W3C Recommendation for VCs, established blockchain protocols for anchoring), portability (DID-based and platform-independent), forgery cost (cryptographic proofs make falsification technically infeasible), and legal deterrence (falsely attributing a third-party attestation carries liability). Their primary weaknesses are high fabrication cost (infrastructure setup, key management) and the absence of GEO evidence: no study has yet tested the effect of cryptographically attested product data on generative engine visibility.

4 Empirical Evidence

While Section 3 characterized each trust signal category individually, this section synthesizes findings across studies to identify four cross-cutting empirical patterns.

Evidence hierarchy and anti-manipulation. Cross-study synthesis reveals a consistent signal hierarchy: evidence-based content (citations, statistics, expert quotes) drives the largest visibility gains in generative engines, structural optimization (section length, content depth) provides moderate benefits, while purely stylistic strategies (tone, keyword stuffing) yield minimal to zero improvement [1, 37, 11, 21]. Notably, generative engines exhibit a natural anti-manipulation property absent in traditional SEO: high keyword semantic matching *reduces* citation frequency, and FAQ schema markup correlates negatively with citations [37], indicating that over-optimization is penalized rather than rewarded.

Source authority and ecosystem divergence. Source authority functions as a cross-cutting modulator of all signal categories. An experimental study of 192,000 assessments [22] demonstrates that source attribution systematically alters LLM content evaluation, even when content quality is held constant, and over 75% of ChatGPT-cited health sources come from established institutional sources [23]. This authority effect is amplified by the divergence between AI and traditional search ecosystems: source overlap is as low as 15% in some product categories [36], 37% of LLM-cited domains are absent from traditional search results entirely [20], and AI engines exhibit a systematic bias toward earned media, with ChatGPT drawing 92.1% of citations from third-party editorial content in consumer electronics while high-traffic domains (10M+ monthly visitors) receive approximately 3× more citations [37, 36]. Together, these patterns create a "big brand advantage": evidence-based signals (Section 3.5) are necessary but not sufficient, as their effectiveness depends on the perceived authority of the content source. For smaller merchants lacking domain authority, portable trust signals that encode verifiable provenance—linking claims to recognized issuers—may serve as an equalizer.

Format versus substance. Structured data provides the machine-readable format that enables AI systems to parse product information, but generative engines evaluate the *substance* conveyed through that format, not its mere presence. Knowledge graphs improve LLM accuracy by 3.4× compared to unstructured data [32], and rich results achieve higher click-through rates [31], yet the counter-intuitive finding that FAQ schema correlates negatively with ChatGPT citations [37] confirms that schema markup makes claims parseable without making them credible. This distinction motivates the verifiable credential approach (Section 6): structured data provides the container; VCs provide the trust.

5 Comparative Analysis

The taxonomy and empirical evidence presented in Sections 3 and 4 enable a structured comparison of the seven trust signal categories across seven evaluation dimensions. Two dimensions deserve clarification: *standardization* measures whether a formal specification exists for the signal format (e.g., a W3C Recommendation or an ISO standard), while *portability* measures whether the signal effectively

functions across discovery platforms in practice. These are related but distinct: a standard can exist without being adopted cross-platform (traditional certifications follow ISO standards but are represented as non-portable images), and a signal can be portable without a formal standard (evidence and citations transfer across platforms through informal conventions). Table 1 summarizes this analysis. This section interprets the table to identify four structural patterns.

The trust signal gap. The central finding of this comparative analysis is that no single trust signal category satisfies all seven evaluation dimensions simultaneously. Categories with the strongest GEO evidence (evidence/citations (e) and social proof (c)) score low on forgery cost and standardization, while platform badges (b), whose evidence is limited to single-platform agent contexts, score high on forgery cost but zero on portability. Conversely, traditional certifications (f) score high on legal deterrence but very low on forgery cost (any merchant can textually claim a certification), while only cryptographic attestations (g) score high on *both* forgery cost and legal deterrence, yet have no direct GEO evidence. This "trust signal gap" constitutes the core structural challenge: among *producible* trust signals, those that generative engines currently reward are precisely those most vulnerable to fabrication, while the only category offering both technical and institutional resistance to manipulation remains untested in AI discovery systems.

The source authority barrier. One factor conspicuously absent from the taxonomy is *source authority* (i.e., domain reputation, organic traffic volume, and earned media presence). As shown in Section 4, it is empirically the single strongest predictor of GEO visibility, yet it is not a *producible signal*: a merchant cannot "add" domain authority to a product listing the way one adds structured data or a certification claim. This creates a paradox: evidence-based signals (e) exhibit the strongest GEO evidence of any category (+30% overall visibility, +93% for data-rich content [1]), yet they operate through an earned media ecosystem that systematically favors established brands, while the categories accessible to all merchants (stylistic optimization (a) and self-declared structured data (d)) offer weak GEO evidence or low forgery cost. For established brands competing against peers of similar authority, the producible signals in the taxonomy become the differentiator. For smaller merchants, the trust signal gap is most acute, and breaking this barrier requires trust signals that convey institutional credibility *without* depending on earned media volume—a role that verifiable third-party attestations (categories f and g) are designed to fill.

The complementarity thesis. The dimension profiles of structured data (d), traditional certifications (f), and cryptographic attestations (g) reveal a natural complementarity. Structured data provides the machine-readable format and high portability that certifications lack, while certifications provide the institutional backing and legal deterrence that structured data lacks. Cryptographic attestations combine both: they inherit the machine-readability and standardization of structured data (JSON-LD serialization, Schema.org alignment) with the third-party attestation model of traditional certifications, adding cryptographic non-repudiation that neither category offers alone. This (d)+(f)→(g) complementarity suggests that cryptographic attestations are architecturally positioned to bridge the trust signal gap, though this theoretical advantage remains empirically unvalidated in GEO contexts.

Portability fragmentation. Cross-platform portability is a structural weakness shared by multiple trust signal categories, not just platform badges. Badges (b) are the most extreme case: within Amazon, they causally increase human conversions (+40.6%) [27] and AI agent selection (+10–33 pp across frontier models) [7, 8], yet they are not exposed as portable metadata and provide zero signal to non-Amazon discovery surfaces. Social proof (c) exhibits a similar fragmentation: Amazon reviews, Google reviews, and Trustpilot ratings each remain siloed on their respective platforms, with no standardized mechanism for aggregating or transferring reputation across ecosystems. Even structured data (d), which is theoretically portable via Schema.org, faces practical fragmentation at the protocol level: ACP requires merchants to push product feeds in JSONL/CSV format to OpenAI's endpoint, while UCP routes data through Google Merchant Center with its own schema [13, 14], forcing merchants to maintain parallel feeds. Given that source overlap between Google and ChatGPT is as low as 15% for some product categories [36], this fragmentation is not a transitional friction but a structural feature of the current landscape. Cryptographic attestations (g) are architecturally positioned to break this pattern: DID-based identifiers and W3C-standardized credential formats function independently of any single platform, enabling a merchant to issue a credential once and have it verified across all discovery surfaces.

6 Cryptographic Attestations: Bridging the Trust Signal Gap

The comparative analysis in Section 5 identified a structural gap: no single trust signal category satisfies all evaluation dimensions, and the categories with the strongest GEO evidence are the most vulnerable to manipulation. This section examines how W3C Verifiable Credentials [10] address this gap, assesses their technical feasibility, and evaluates their integration prospects with emerging agentic commerce protocols.

Table 1: Comparative evaluation of trust signal categories across seven dimensions. Ratings synthesize evidence from the studies reviewed in Sections 3 and 4.

Signal Category	Prod. Ease	GEO Evid.	Machine-Read.	Standard.	Portab.	Forgery Cost	Legal Deter.
(a) Stylistic	●	○[a]	◔	—	—	○	○
(b) Platform badges	◕[b]	◑	◑	○	—	◕	◔
(c) Social proof	◔	◕[c]	◑	◑	◑	◔	○
(d) Structured data	◕	◑[d]	●	◕	◕	○	○
(e) Evidence/citations	◔	●	◑	—	◑	○	◔
(f) Trad. certifications	○	—[e]	◔	◑	◔	○	◕
(g) Crypto. attestations	○	—[e]	●	◕	●	●	◕

● Very High ◕ High ◑ Medium ◔ Low ○ Very Low — N/A

[a] Mixed results across studies. [b] Vendor-side cost. [c] Aggregated ratings only. [d] Growing evidence base. [e] Untested in GEO contexts.

6.1 Mapping VCs to the Taxonomy

VCs intersect multiple taxonomy categories simultaneously, functioning as a composite signal rather than a standalone category. As *structured data enrichment*, VCs serialized in JSON-LD are compatible with the Schema.org markup that AI shopping agents rely on for product indexing [19]. The UN Transparency Protocol (UNTP), adopted in July 2025 as UN/CEFACT Recommendation No. 49 [15], mandates W3C VCDM 2.0 as the credential format for Digital Product Passports, conformity credentials, and traceability events. Combined with GS1's VC/DID technical landscape [17], these standards avoid the fragmentation of proprietary certification formats. Digital Product Passport (DPP) implementations demonstrate that DIDs and VCs can manage product lifecycle data across supply chains [16].

As *verifiable evidence*, each credential constitutes a citable, dated claim with a verification mechanism, structurally analogous to the citation-heavy content that produces the largest GEO visibility gains (+30% to +115%) [1]. A product listing enriched with multiple VCs (material composition from a testing lab, origin from the manufacturer, sustainability from an auditor) embeds multiple evidence-backed claims from distinct authoritative sources.

6.2 Technical Feasibility

Credential verification is not a task for LLMs themselves but for the platforms that orchestrate them. In production systems, agentic commerce platforms (Google, OpenAI, Perplexity) would integrate VC verification into their indexing and discovery pipelines—resolving issuer DIDs, checking revocation status, and validating signatures—before surfacing a structured trust assessment to the LLM. The LLM's role is then limited to interpreting the *content* of the VC and the verification outcome, not performing cryptographic operations.

The broader deployment landscape constrains near-term expectations. A study of nine real-world decentralized identity deployments [35] finds high attrition among early systems. Government-backed initiatives dominate (6 of 9 deployments), and the five-layer interoperability challenge (key management, DID methods, credential formats, exchange protocols, trust frameworks) [33] remains largely unresolved.

6.3 Integration with Agentic Commerce Protocols

The two major agentic commerce protocols exhibit varying degrees of VC compatibility. UCP's interoperability layer supports Agent2Agent (A2A) and Model Context Protocol (MCP) [14], providing natural extension points through which platforms could surface verified credential assessments. ACP's structured conversation model [13] could accommodate VC presentation during the discovery and evaluation phases of buyer-merchant interaction.

However, practical barriers remain substantial. The SSI ecosystem's adoption trajectory (characterized by high attrition, government dependence, and limited scale [35]) suggests that merchant-side VC infrastructure will not emerge organically. Finally, the agentic commerce ecosystem itself is evolving rapidly, with ACP and UCP both in early stages with specifications subject to change.

7 Open Challenges and Research Agenda

The preceding analysis reveals six open challenges that define a research agenda for trust signals in AI-mediated commerce.

Challenge 1: Empirical validation of cryptographic attestations in GEO. The central gap identified by the comparative analysis (Section 5) is that the trust signal category scoring highest on forgery cost, machine-readability, standardization, and portability (cryptographic attestations (g)) has *no* GEO evidence. No study has measured whether cryptographically attested product data—whether via VCs, on-chain anchoring, or signed metadata—affects generative

engine visibility, citation likelihood, or AI agent selection rates. Establishing this link requires controlled experiments that embed cryptographic attestations in product listings and measure differential treatment by generative engines, while controlling for position bias [25] and the confounding effect of content quality. Until this evidence exists, the theoretical advantages of cryptographic attestations remain unvalidated in the context that matters most: actual AI discovery and recommendation.

Challenge 2: Platform-side credential verification. For cryptographic attestations to function as effective trust signals, agentic commerce platforms (Google, OpenAI, Perplexity) must integrate credential verification into their discovery pipelines. Without platform-side verification, the market faces a classic *lemon problem* [24]: if AI engines cannot distinguish verified from self-declared claims, merchants have no incentive to invest in obtaining legitimate credentials, and low-quality signals crowd out high-quality ones. The technical components—DID resolution, signature validation, revocation checking—are well-understood; the open question is whether and when major platforms will adopt them. This is fundamentally an adoption and incentive challenge, not a research problem, but it conditions the practical impact of every other trust signal category examined in this paper.

Challenge 3: Interoperability and data fragmentation. The trust signal landscape suffers from multi-layered fragmentation. At the *protocol level*, ACP and UCP define incompatible product data pipelines: ACP requires merchants to push product feeds in JSONL/CSV format to OpenAI's centralized endpoint (with platform-specific fields such as `is_eligible_search`), while UCP routes product data through Google Merchant Center with its own feed format and a `native_commerce` flag [13, 14]. A merchant wishing to be discoverable on both ChatGPT and Google must maintain two separate feeds with different schemas. At the *structured data level*, Schema.org markup on merchant websites follows yet another format (JSON-LD microdata), with no guaranteed ingestion by either protocol. At the *credential level*, UNTP mandates W3C VCDM 2.0 for Digital Product Passports [15], but neither ACP nor UCP currently defines how VCs should be embedded in or alongside product feeds. The DID/VC ecosystem itself faces interoperability challenges spanning five layers (key management, DID methods, credential formats, exchange protocols, and trust frameworks) with identified security threats requiring coordinated mitigation [33]. At the *trust anchoring level*, current VC verification relies on the issuer's own infrastructure for DID resolution and revocation checking, creating a centralization of trust; lightweight distributed anchoring mechanisms (e.g., `did:webvh` with verifiable history) could provide the auditability and persistence guarantees needed for long-term credential verification, particularly in post-sale scenarios where the original issuer may have ceased operations.

Challenge 4: Big-brand bias and market concentration. Empirical evidence reveals a systematic bias toward established brands in AI search: ChatGPT draws 92.1% of citations from earned media in consumer electronics [36], sites with 10M+ organic traffic receive 3× more citations [37], and source overlap between platforms is as low as 15% [36]. This "big brand advantage" risks reinforcing market concentration, as smaller merchants cannot easily generate the earned media presence that drives AI visibility. Trust signals that are portable, standardized, and independent of existing brand authority (such as VCs issued by recognized certification bodies) could serve as an equalizer, but only if generative engines learn to weight them. Research is needed on whether verifiable third-party attestations can substitute for domain authority as a credibility signal in AI-mediated discovery.

Challenge 5: Regulatory momentum and adoption trajectory. Two regulatory developments are poised to resolve the adoption question that has historically constrained decentralized identity deployments [35]. The EU Digital Product Passport (DPP) regulation, effective from 2027 for priority product categories, mandates machine-readable product data that aligns naturally with VC-based attestations [15, 16]. The eIDAS 2.0 regulation establishes the European Digital Identity Wallet framework, creating government-backed infrastructure for VC issuance and verification. The adoption of UNTP as UN/CEFACT Recommendation No. 49 [15] further standardizes VCs as the prescribed credential format for supply-chain transparency worldwide. Together, these create both supply-side pressure (merchants must provide structured, verifiable product data) and infrastructure availability (identity wallets and trust registries). The remaining research questions concern the *transition period*: how quickly issuers and verifiers will reach operational readiness before regulatory deadlines, how compliance-driven VC issuance interacts with the voluntary trust signal landscape, and whether markets outside the EU regulatory perimeter will adopt UNTP voluntarily or require analogous mandates.

Challenge 6: Post-sale lifecycle and secondary markets. UNTP's primary operational scope is *Cradle-to-Gate*: it standardizes how manufacturers, suppliers, and conformity bodies produce and exchange sustainability data up to the point of first sale [15]. Yet the secondary market (resale, repair, refurbishment) is growing rapidly: global recommerce reached $188 billion in 2024 and is projected to surpass $310 billion by 2029, expanding roughly 3× faster than primary retail [38]. AI agents will increasingly recommend

pre-owned products alongside new ones. Post-sale lifecycle events (ownership transfers, maintenance records, condition assessments) fall outside UNTP's core scope and lack a standardized trust framework. Once a product leaves the manufacturer's supply chain, no single institutional authority exists to anchor trust. This authority vacuum raises the question of how post-sale trust signals should be represented and verified so that AI agents can recommend second-hand products with the same confidence as new ones, and whether decentralized registries are needed precisely because no central authority can fill this role. Extending VC-based traceability to consumer-to-consumer transfers represents a concrete research direction with direct implications for circular economy and AI-mediated discovery.

8 Conclusion

This systematization has examined the trust signal landscape in AI-mediated commerce through a seven-category taxonomy evaluated across seven dimensions. The central finding is a *trust signal gap*: the signals with the strongest generative engine evidence (evidence/citations (e), social proof (c), and platform badges (b)) score low on forgery cost or portability, while only cryptographic attestations (g) combine both technical and institutional resistance to manipulation, yet have no direct GEO evidence. Among cryptographic attestation approaches, W3C Verifiable Credentials emerge as the most mature and standardized candidate to bridge this gap, combining the machine-readability of structured data (d) with the institutional attestation model of certifications (f) and adding cryptographic non-repudiation.

The most urgent research priority is empirical validation: until controlled experiments demonstrate that generative engines differentially treat cryptographically attested content, the case for this category as a GEO signal remains analytical rather than empirical. Two regulatory drivers may accelerate progress: the EU Digital Product Passport regulation, mandating machine-readable product attestations from 2027 [15, 16], and the eIDAS 2.0 framework providing government-backed VC infrastructure. This systematization provides a structured framework for evaluating trust signals as they emerge, and identifies the dimensions along which future solutions must improve to close the trust signal gap.

References

[1] P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, and A. Deshpande, "GEO: Generative Engine Optimization," in *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2024. DOI: 10.1145/3637528.3671900.

[2] OpenAI, "Introducing Shopping Research in ChatGPT," 2025.

[3] Google, "New Technology for Retailers in the Agentic Shopping Era," 2026.

[4] Perplexity, "Shop Like a Pro," 2025.

[5] "Trust at Risk: Detecting Misinformation in LLM-Generated Product Reviews," *ScienceDirect*, 2025. DOI: S2772503025000994.

[6] "Large Language Models as 'Hidden Persuaders': Fake Product Reviews are Indistinguishable to Humans and Machines," 2025. arXiv: 2506.13313.

[7] A. Allouah *et al.*, "What Is Your AI Agent Buying? Evaluation, Biases, Model Dependence, and Emerging Implications for Agentic E-Commerce," Columbia Business School, 2025. arXiv: 2508.02630.

[8] I. Cherep, S. Ma, H. Xu, Y. Shaked, P. Maes, and A. Singh, "A Framework for Studying AI Agent Behavior: Evidence from Consumer Choice Experiments," in *NeurIPS 2025 Workshop on LLM Agents (LAW)*, 2025. arXiv: 2509.25609.

[9] W3C, "Decentralized Identifiers (DIDs) v1.0," W3C Recommendation, July 2022. https://www.w3.org/TR/did-core/

[10] W3C, "Verifiable Credentials Data Model v2.0," W3C Recommendation, May 2025.

[11] P. S. Bagga, V. F. Farias, T. Korkotashvili, T. Peng, and Y. Wu, "E-GEO: A Testbed for Generative Engine Optimization in E-Commerce," 2025. arXiv: 2511.20867.

[12] "CORE: Controlling Output Rankings in Generative Engines for LLM-based Search," 2026. arXiv: 2602.03608.

[13] Stripe, "Developing an Open Standard for Agentic Commerce," 2025.

[14] Google, "Under the Hood: Universal Commerce Protocol (UCP)," 2026.

[15] UN/CEFACT, "UN Transparency Protocol (UNTP)," Recommendation No. 49, adopted July 2025. https://untp.unece.org/

[16] "Digital Product Passport Management with Decentralised Identifiers and Verifiable Credentials," 2024. arXiv: 2410.15758.

[17] GS1, "VCs and DIDs Technical Landscape," 2025.

[18] Wellows, "Google AI Overviews Ranking Factors 2026," 2026.

[19] Google, "Product data specification — Google Merchant Center Help," 2025.

[20] P. Zhang, Q. Ye, Z. Peng, K. Garimella, and G. Tyson, "Source Coverage and Citation Bias in LLM-based vs. Traditional Search Engines," 2025. arXiv: 2512.09483.

[21] K.-C. Yang, "News Source Citing Patterns in AI Search Systems," 2025. arXiv: 2507.05301.

[22] F. Germani and G. Spitale, "Source Framing Triggers Systematic Bias in Large Language Models," *Science Advances*, 2025. DOI: 10.1126/sciadv.adz2924.

[23] E. Jacques *et al.*, "Authority Signals in AI Cited Health Sources: A Framework for Evaluating Source Credibility in ChatGPT Responses," 2026. arXiv: 2601.17109.

[24] G. A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488–500, 1970.

[25] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," *Trans. Assoc. Comput. Linguistics*, vol. 12, 2024. DOI: 10.1162/tacl_a_00638.

[26] G. Filandrianos, A. Dimitriou, M. Lymperaiou, K. Thomas, and G. Stamou, "Bias Beware: The Impact of Cognitive Biases on LLM-Driven Product Recommendations," in *Proc. EMNLP 2025*, 2025. arXiv: 2502.01349.

[27] M. Bairathi, X. Zhang, and A. Lambrecht, "The Value of Platform Endorsement," *Marketing Science*, vol. 44, no. 1, pp. 84–101, 2025. DOI: 10.1287/mksc.2022.0226.

[28] X. Zhao, W. You, Z. Zheng, S. Shi, Y. Lu, and L. Sun, "How Do Consumers Trust and Accept AI Agents? An Extended Theoretical Framework and Empirical Evidence," *Behavioral Sciences*, vol. 15, no. 3, 337, 2025. DOI: 10.3390/bs15030337.

[29] Visa Intelligent Commerce, "Earning Consumer Trust in the Age of Agentic Commerce: Understanding How AI Assistants Are Reshaping the Buyer Journey," Visa Agentic Commerce Consumer Research Surveys, 2025.

[30] A. Volpini and N. Demir, "Structured Data," in *Web Almanac 2024*, HTTP Archive, November 2024.

[31] Milestone Research, "Rich Results CTR Study," 2020. See also: M. G. Southern, "Google SERP Study: Which Rich Results Get the Most Clicks?," *Search Engine Journal*, 2020.

[32] J. Sequeda, D. Allemang, and B. Jacob, "A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases," in *Proc. 7th GRADES-NDA Workshop*, ACM SIGMOD, 2024. DOI: 10.1145/3661304.3661901.

[33] C. Mazzocca, A. Acar, S. Uluagac, R. Montanari, P. Bellavista, and M. Conti, "A Survey on Decentralized Identifiers and Verifiable Credentials," *IEEE Communications Surveys & Tutorials*, 2025. DOI: 10.1109/COMST.2025.3543197.

[34] S. R. Garzon, A. Vaziry, E. M. Kuzu, D. E. Gehrmann, B. Varkan, A. Gaballa, and A. Küpper, "AI Agents with Decentralized Identifiers and Verifiable Credentials," in *Proc. 18th Int. Conf. Agents and Artificial Intelligence (ICAART)*, 2026. arXiv: 2511.02841.

[35] D. Schumm, K. O. E. Müller, and B. Stiller, "Are We There Yet? A Study of Decentralized Identity Applications," University of Zurich, 2025. arXiv: 2503.15964.

[36] M. Chen, X. Wang, K. Chen, and N. Koudas, "Generative Engine Optimization: How to Dominate AI Search," 2025. arXiv: 2509.08919.

[37] SE Ranking Research / Higglo Digital, "The Real Factors Behind ChatGPT Citations: What Our Study of 129,000 Domains Reveals," 2025.

[38] "Recommerce Intelligence Report 2025: Market to Surpass $310 Billion by 2029," GlobeNewsWire / Research and Markets, July 2025.