Journal of Digital Content Management (JDCM)
Articles in Press
dcm.atu.ac.ir
DOI:

ATU PRESS

Original Research
Accepted: --------
Received: ------
eISSN: 3116-1448
ISSN: 3116-1715

Generating Trustworthy Structured Data with Large Language Models: The Role and Limitations of Schema.org

Mehdi Taheri *

Department of Knowledge & Information Science , Allameh Tabataba'i University, Tehran, Iran

Abstract

The growing use of large language models (LLMs) in information services has increased demand for outputs that can be processed, linked, validated, and reused by digital content management systems. Yet, machine-readable form does not ensure semantic coherence, factual accuracy, or traceability. This study analyzes challenges in LLM-based structured data generation and examines the contribution and limits of Schema.org. Using a qualitative, descriptive-analytical design and directed documentary content analysis, it organizes the challenges into five dimensions: generation and stability; structure and conformance; semantics and identity; accuracy and evidence; and provenance and traceability. The analysis shows that Schema.org can support shared semantic description through common entity types, properties, and relationships, improving consistency, interoperability, and reuse. It cannot independently validate syntax, guarantee factual accuracy, resolve entity identity, verify relationships, or establish claim provenance. These functions require complementary mechanisms: JSON-LD or RDF for representation; JSON Schema or SHACL for constraint validation; authority data, persistent identifiers, and knowledge graphs for identity and relationship management; and evidence-based retrieval and provenance models for verification and traceability. The study proposes a layered framework in which trustworthiness is assessed through five related dimensions: syntactic validity, structural and

semantic conformity, factual accuracy and evidential support, provenance and traceability, and operational reusability. The framework positions Schema.org as a semantic component within a broader digital content management architecture.

Keywords: Large Language Models; Structured Data Generation; Data Trustworthiness; Schema.org; Digital cContent Management

Introduction

Large language models (LLMs) and generative chatbots have changed how people access and use digital information. Whereas conventional information retrieval typically presents documents, web pages, or metadata records for users to inspect, generative systems synthesize information into direct responses. This shift can make information access more convenient, but it also changes where errors arise and how they propagate. A fluent answer may be persuasive while containing an inaccurate name, date, identifier, attribute, or relationship. When such an answer is transferred into a repository, content management system, or automated workflow, an error in a conversational response can become an error embedded in an information infrastructure (Augenstein et al., 2024; Ji et al., 2023; Wang et al., 2024).

This problem becomes more consequential when a chatbot is expected to produce information that is precise, organized, machine-processable, and reusable. A description of a scholarly work, for example, may need to identify its title, creators, publication date, publisher, identifiers, and relationships to other works. A digital library or research repository must be able to interpret these elements consistently and, where necessary, reconcile them with authority records and other sources. Similar requirements arise in archives, museums, publishing, e-commerce, and public-sector information systems. In these environments, readability and relevance are insufficient measures of output quality; the data must also be interpretable, verifiable, and suitable for downstream use.

The term structured data covers representations organized according to an explicit structure or schema. Examples include relational records, JSON objects with defined keys and values, and graph representations consisting of entities and relationships. Yet the presence of a structure does not establish that its meaning is shared or its content is correct. At least three levels must therefore be distinguished: syntactic validity, which concerns conformance to a serialization format; structural and semantic conformity, which concerns the appropriate use and interpretation of types, properties, values, and relationships; and factual validity, which concerns whether represented claims are supported by reliable evidence. A syntactically valid JSON object may contain a fabricated publication date, while a factually correct value may be assigned to the wrong entity or property.

Articles in Press

Hallucination is one prominent manifestation of this problem. Research has documented the variety of unsupported or factually inconsistent content generated by language models and the difficulty of evaluating factuality across contexts (Huang et al., 2023; Ji et al., 2023). In structured outputs, such errors can appear as fabricated values, invalid identifiers, mistaken entity matches, or relationships that have no evidential basis. The problem is not limited to whether a model can produce JSON or a table. It concerns whether generated content can be transformed into data whose meaning, identity, relationships, and evidential status remain assessable after the initial interaction.

Semantic standardization is relevant because information systems often use different vocabularies and structures to describe similar entities. Shared vocabularies can reduce variation in the naming and interpretation of information elements and facilitate data exchange. Schema.org is a prominent Web vocabulary for describing entities, properties, and relationships (Taheri, et al. 2018). It includes types such as Person, Organization, Book, Article, Event, and Product, and can be expressed using syntaxes including JSON-LD, Microdata, and RDFa (Schema.org, n.d.-a, n.d.-b). Its role in generative systems, however, needs to be specified carefully. A shared vocabulary can guide semantic mapping and improve the consistency of descriptions; it cannot determine whether a generated claim is true, whether two names refer to the same entity, or whether a relationship is supported by evidence.

This study examines the relationship between LLM-based structured data generation and semantic standardization. It asks what challenges affect the trustworthiness of generated structured data, which of these challenges Schema.org can mitigate, and what complementary mechanisms are required across the digital content management lifecycle. The central argument is that trustworthiness emerges from coordinated semantic, structural, evidential, identity-management, and provenance processes. Schema.org contributes to this architecture as a shared semantic vocabulary, while other technologies and governance practices perform distinct functions that should not be conflated with vocabulary-based description.

Literature Review

LLMs are trained to model statistical regularities in language and generate sequences conditioned on input context. Transformer-based architectures have enabled substantial advances in language processing

and generation (Vaswani et al., 2017). These capabilities support summarization, extraction, classification, question answering, and the production of structured-looking outputs. However, a model's generative objective does not make it an authoritative database or a system of record. Its output may reflect incomplete training data, ambiguous prompts, context loss, or probabilistic generation. Work on factuality and trustworthiness has consequently emphasized the need to assess claims against evidence rather than infer reliability from fluency or plausibility (Augenstein et al., 2024; Wang et al., 2024).

The distinction between natural-language generation and structured data generation is important. In a conversational answer, a human reader may infer the intended relationship between a work and its author from sentence context. A receiving information system requires that the entities be identified, the relationship be represented in an agreed form, and the values be mapped to appropriate properties. The conversion process therefore involves several operations: recognizing entities, extracting values, disambiguating references, selecting semantic types and properties, representing the result in a machine-readable syntax, validating constraints, and checking claims against sources. Errors at one stage can remain hidden when later stages produce a well-formed representation.

The Semantic Web and Linked Data approaches provide a technical and conceptual basis for explicit, linkable descriptions of resources. Linked Data emphasizes the use of identifiers, structured descriptions, and links among resources, enabling information to be connected across systems (Berners-Lee, 2006; Bizer et al., 2009; Heath & Bizer, 2011). RDF represents information as graph-based statements, while JSON-LD provides a JSON-based syntax for expressing linked data (Cyganiak et al., 2014; W3C, 2020). These technologies should be distinguished from vocabularies: a serialization syntax determines how data are expressed, a graph data model determines how statements are structured, and a vocabulary supplies terms used to describe the things and relationships represented.

Schema.org provides a widely adopted vocabulary intended to make structured descriptions of Web content more interoperable. Its types and properties offer a common descriptive resource for publishers and data consumers. The vocabulary is intentionally broad and flexible, and its data model supports extension and varied implementation contexts (Schema.org, n.d.-a, n.d.-b). That flexibility is useful for broad

Articles in Press

Web publishing, but it should not be mistaken for a complete constraint language or a domain-specific application profile. The existence of a property does not make it mandatory in every context, nor does the vocabulary alone establish cardinality, completeness, or all local data-quality requirements.

Validation technologies address a different problem. JSON Schema can specify constraints for JSON documents, while the Shapes Constraint Language (SHACL) can express and validate constraints over RDF graphs (JSON Schema, 2020; W3C, 2017). Such mechanisms can check whether data meet declared structural or graph constraints, but validation against a schema does not independently prove that a value is factually correct. A record may pass validation while containing a false date or an incorrectly identified person. This distinction between conformance and truth is central to assessing generated data.

Entity identification and linking introduce another set of requirements. Entity recognition identifies a mention or candidate entity in text; entity linking attempts to associate that mention with a specific entity in a knowledge base or authority source. Similar names, aliases, transliteration variants, and incomplete context can lead to mistaken matches. Research on entity linking has examined these issues and the methods used to resolve them (Shen et al., 2015). Schema.org properties such as identifier and sameAs can represent identifiers or links, but they do not replace the reconciliation process needed to establish identity. In particular, identifier and sameAs have different semantics and should not be treated as interchangeable.

Knowledge graphs can support the management and retrieval of entities and their relationships. They make explicit connections available for querying and can provide external context to language models. Research on knowledge graphs and LLM integration identifies opportunities for grounding, retrieval, and reasoning, while also noting that the quality of the graph and the model's use of retrieved information remain decisive (Hogan et al., 2021; Lavrinovics et al., 2025; Pan et al., 2024). Retrieval-augmented generation (RAG) similarly enables models to use information retrieved from external sources, but retrieval does not guarantee correctness if sources are unreliable, retrieval is incomplete, or the model misinterprets evidence (Lewis et al., 2020).

The literature thus points to a need for an integrated account of trustworthiness. Semantic vocabularies, serialization formats,

validation languages, authority data, knowledge graphs, retrieval systems, and provenance models have complementary roles. Treating any one of them as a complete solution obscures the different sources of error and the controls required to address them. The present study brings these functions together in an analytical framework focused on the transformation of LLM outputs into structured data suitable for digital content environments.

Purpose of the Study

The purpose of this study is to analyze the challenges affecting the generation of trustworthy structured data by LLMs and to determine the contribution and limits of Schema.org in addressing them. It addresses three questions: (1) What categories of challenge affect the generation, structure, semantics, accuracy, identity, and traceability of LLM-generated data? (2) Which challenges can be mitigated through Schema.org, and which require complementary technical or organizational mechanisms? (3) How can Schema.org be positioned within a broader architecture for validating, managing, retrieving, and reusing generated data across the digital content management lifecycle?

The study uses trustworthiness as an analytical construct with five evaluation dimensions: syntactic validity; structural and semantic conformity; factual accuracy and evidential support; provenance and traceability; and operational reusability. These dimensions are related but not interchangeable. A successful outcome in one dimension does not establish success in the others.

The five dimensions also clarify the analytical boundary between data quality and trustworthiness. Data quality is often operationalized through properties such as accuracy, completeness, consistency, and timeliness. In LLM-generated data, these properties need to be assessed in relation to the process that produced the record and the evidence available for its claims. Completeness, for example, should not be pursued by filling every empty field with a plausible value. An explicit unknown value may be more trustworthy than an unsupported completion. Likewise, consistency across fields is not sufficient if the fields consistently reproduce an incorrect source. The framework therefore treats evidence and provenance as constitutive parts of assessment rather than as optional documentation added after data generation.

Articles in Press

Method

The study adopts a qualitative, descriptive-analytical design and uses documentary research with directed qualitative content analysis. Its purpose is conceptual and analytical: to distinguish the functions performed by language models, semantic vocabularies, representation formats, validation mechanisms, identity services, and provenance models in the production and management of structured data. The study does not test the performance of a particular model or system.

The documentary materials comprise three related groups: (1) scholarly studies addressing large language models, hallucination, factuality, output stability, and trustworthiness; (2) scholarly and technical sources on structured data, Linked Data, RDF, JSON-LD, entity linking, knowledge graphs, and constraint validation; and (3) the official Schema.org documentation and relevant W3C and JSON Schema specifications. Sources are selected purposively for their direct relevance to the research questions, their scholarly or technical authority, and their contribution to defining concepts, capabilities, constraints, or evaluation criteria. The analysis gives priority to peer-reviewed research and primary technical specifications when establishing the functions of a technology or standard.

The documentary analysis is directed rather than exhaustive or protocol-driven. It is not presented as a systematic review or a meta-analysis; accordingly, the study makes no claim that the documentary corpus represents every publication on LLM trustworthiness or structured data generation. Nor does it report database-wide retrieval counts or a PRISMA screening flow. Its evidential scope is the body of scholarly and technical sources cited in the article and used to support the conceptual comparison. This boundary is important when interpreting the findings: the categories are an analytical synthesis, not a frequency ranking of problems in all LLM systems.

The analysis proceeds in four stages. First, reported challenges are identified and compared in terms of where they affect the production and subsequent use of structured data. Second, Schema.org is examined through its types, properties, relationships, and documented data-model principles to determine what it can represent and what it does not formally enforce. Third, each challenge is mapped against Schema.org's direct or indirect contribution and against complementary mechanisms required where the vocabulary is insufficient. Fourth, the results are synthesized into a layered framework and assessed against five

dimensions of trustworthiness: syntactic validity; structural and semantic conformity; factual accuracy and evidential support; provenance and traceability; and operational reusability.

The resulting classification comprises 18 specific challenges organized into five analytical categories: generation and stability; structure and conformance; semantics and identity; accuracy and evidence; and provenance and traceability. Trustworthiness is treated as the overall quality construct to be evaluated, rather than as an additional challenge at the same analytical level. This distinction separates individual failure modes from the multidimensional quality they may affect.

Results

The documentary analysis identifies five interrelated categories of challenge: generation and stability; structure and conformance; semantics and identity; accuracy and evidence; and provenance and traceability. The categories locate the principal failure modes while recognizing that a single output may exhibit several at once. For example, an unstable response may contain a different identifier, which can create an identity error and weaken provenance. Table 1 summarizes the 18 challenges, their manifestations, Schema.org's potential contribution, and the complementary controls required.

The classification distinguishes between problems that concern the model's generative behavior and those that arise when generated content is converted into managed data. Generation-related challenges include non-determinism, response instability, repeatability, and prompt sensitivity. Structural challenges concern the conversion of text into a constrained representation and the conformity of that representation to an agreed vocabulary or application profile. Semantic and identity challenges concern the meanings assigned to properties, the identification of entities, and the correctness of relationships. Accuracy and evidence challenges concern whether claims are true, current, and supported. Provenance challenges concern whether users can identify where a value came from, when it was retrieved, and how it was produced or verified.

Schema.org's contribution is strongest where the problem concerns the shared description of entities, properties, and relationships. It can reduce variation in terminology, provide a common semantic reference, and make generated descriptions more readily interpretable by systems

Articles in Press

that use the vocabulary. It can also support the expression of links to other resources. Its contribution is indirect or limited where the problem concerns syntax, application-specific completeness, factual correctness, entity reconciliation, evidence quality, or process reproducibility. Those tasks require additional controls.

Previous research on the localization of Schema.org for manuscript description has demonstrated that the vocabulary can be extended and adapted to accommodate domain-specific descriptive requirements. By examining the ontological structure of Schema.org, including the inheritance of properties from parent types and the definition of additional properties for manuscript description, Mohammadi Ostani et al. (2021) showed how the vocabulary could be enriched to represent contextual characteristics that are not adequately covered by its general-purpose terms. This finding has an important implication for LLM-generated structured data: the semantic adequacy of an output depends not only on selecting an existing type or property, but also on whether the vocabulary provides sufficient conceptual coverage for the domain being described. Accordingly, the use of Schema.org in specialized information environments may require domain-specific extensions and carefully defined application profiles.

Table 1. Classification of challenges in LLM-based structured data generation and the role of Schema.org

Category	**Challenge**	**Manifestation in generated data**	**Schema.org contribution**	**Complementary mechanisms**	**Trustworthiness implication**
Generation and stability	Probabilistic and non-deterministic outputs	Different values or relationships for the same entity across similar prompts.	Shared types and properties stabilize description, not model behavior or truth.	Controlled prompts; source retrieval; post-generation checks.	Improves consistency of representation, not factual accuracy.
Generation and stability	Response instability across similar prompts	Changes in values, field order, naming, or record detail.	Common terms reduce variation in labels and property selection.	Fixed output templates; generation controls; regression tests.	Improves comparability without ensuring identical outputs.
Generation and stability	Limited repeatability and reproducibility	Outputs vary across runs, model versions, or settings.	A shared vocabulary supports comparable representations.	Record model/version, prompts, parameters, sources, timestamps, and transformations.	Makes processes auditable and repeatable.

Generation and stability	Prompt sensitivity	Small changes in prompt wording alter values or structure.	Vocabulary terms provide stable semantic targets.	Standardized prompts; explicit field rules; missing-value policies.	Reduces representational variation, not prompt sensitivity itself.
Structure and conformance	Text generation instead of data generation	Explanatory prose, missing fields, or information in unsuitable fields.	Types and properties provide a target structure for mapping.	Constrained generation; JSON-LD templates; automated parsing.	Improves conversion into machine-processable records.
Structure and conformance	Lack of structural guarantees	Malformed JSON, incomplete fields, or inappropriate data types.	Defines semantic terms, not serialization correctness.	JSON parsing; JSON Schema; SHACL for RDF graphs.	Reduces structural errors when constraints are explicit.
Structure and conformance	Nonconformance with a vocabulary	Renamed properties, undefined terms, or incorrect type selection.	Directly supplies shared types and property terms.	Vocabulary checking; application profiles; validation rules.	Improves semantic conformity.
Structure and conformance	Incompatibility with target standards	Different date formats, field structures, or value conventions.	Provides shared descriptive terms but not complete system compatibility.	Transformation rules; application profiles; target-system validation.	Facilitates exchange when mappings are explicit.
Structure and conformance	Absence of comprehensive validation	Plausible records pass unchecked despite violating requirements.	Can inform expected meanings but is not a validation engine.	JSON Schema or SHACL; business rules; source verification.	Improves detection of structural and semantic defects.
Semantics and identity	Difficulty maintaining semantic consistency	Contradictory values, types, or contextual interpretations.	Shared definitions support consistent use of concepts.	Domain rules; cross-field checks; semantic review.	Reduces inconsistency when governance rules are applied.
Semantics and identity	Incorrect entity relationships	Misattribution of authorship, affiliation, membership, or other links.	Relational properties can express the intended relationship.	Authority sources; persistent identifiers; graph-based relationship checks.	Improves explicitness and reviewability of relationships.
Semantics and identity	Entity ambiguity	Different people, works, or organizations have similar names.	Types and properties describe entity roles but do not resolve identity.	Contextual entity resolution; authority	Improves identification when external evidence is used.

Articles in Press

				control; reconciliation.	
Semantics and identity	Invalid or duplicate identifiers	Fabricated, malformed, duplicated, or misassigned identifiers.	identifier and sameAs can represent identifier values and identity links.	Resolve identifiers against authoritative registries; validate URIs and identity assertions.	Improves linking only when identifiers are verified.
Accuracy and evidence	Hallucinated claims	Plausible statements conflict with facts or available evidence.	Provides a place to represent claims, not a method to verify them.	RAG; source evaluation; claim-to-evidence checking; human review.	Raises verifiability and reduces unsupported acceptance.
Accuracy and evidence	Fabricated values	Invented names, dates, URLs, identifiers, or organizational details.	Standardizes representation but cannot verify values.	Authoritative databases; field-level evidence; identifier resolution.	Reduces the entry of false values into repositories.
Accuracy and evidence	Outdated or incomplete model knowledge	Stale information about people, organizations, standards, or products.	Can represent current values and links but does not update them.	Retrieve current sources; record retrieval dates; check validity periods.	Improves currency and evidence assessment.
Accuracy and evidence	Failure to distinguish facts from inferences	Extracted, inferred, and guessed values appear equally certain.	Can describe values and source links but does not assign epistemic status automatically.	Label evidence status; record claim-level provenance; preserve unknowns.	Makes evidential strength visible to downstream users.
Provenance and traceability	Unclear information provenance	Source, retrieval time, extraction method, or verification status is unknown.	Can express some source-related links but is not a complete provenance model.	Record source per claim; timestamps; process metadata; provenance vocabularies.	Improves transparency, auditability, and traceability.

Note. The five categories organize 18 specific challenges. Trustworthiness is treated as the overall outcome assessed across the framework, not as an additional challenge category.

Schema.org and the Boundaries of Its Contribution

Schema.org offers a common semantic vocabulary for describing entities and their properties. For an LLM-generated description of a book, for example, a system can map the title, creator, publisher, publication date, and identifier to properties associated with an

appropriate type. This makes the intended role of each value more explicit than an unstructured sentence and supports processing by systems that understand the same vocabulary. The benefit is semantic coordination: producers and consumers can refer to shared terms rather than relying entirely on locally invented field names.

The vocabulary can also support the representation of relationships among entities. Properties such as author, publisher, member, and affiliation allow a description to connect a resource with people or organizations. These relationships are valuable in information environments where content is not managed as isolated records but as interconnected entities. Nevertheless, a property that expresses a relationship does not establish that the relationship is true. A model may assign a work to the wrong author and still produce a description that uses a recognized property correctly. Relationship representation must therefore be separated from relationship verification.

A related distinction concerns entity identity. Schema.org supports the description of entities and provides properties such as identifier and sameAs, but these terms serve different purposes. An identifier value describes an identifier associated with an item; sameAs expresses identity between the described resource and another resource. Neither property independently establishes that a generated identifier is valid or that two resources are identical. An incorrect sameAs assertion can merge distinct entities and propagate errors across systems. Identity claims should be based on authoritative identifiers, contextual evidence, and explicit reconciliation procedures (Shen et al., 2015).

Schema.org should also be distinguished from the formats and constraint languages used with it. JSON-LD is a syntax for expressing linked data in JSON; RDF provides a graph data model; Schema.org supplies terms that can be used in such representations. JSON Schema can validate the structure of a JSON document, while SHACL can validate constraints over RDF graphs. A record can conform to a JSON Schema or SHACL shape and still contain a false claim. Conversely, an accurate value may be unusable if its representation is malformed or its semantic role is unclear. A robust system therefore requires separate checks for syntax, vocabulary use, application-specific constraints, and factual support.

The flexibility of Schema.org is a design characteristic, not a defect. A broad vocabulary used across many Web contexts cannot be expected to encode every local requirement for completeness,

Articles in Press

cardinality, controlled values, or workflow. Libraries, archives, museums, publishers, and government agencies may require additional elements, authority control, or constraints that are specific to their resources and policies. Application profiles and domain rules are therefore necessary when a general vocabulary is adopted in a specialized information environment. The vocabulary can provide a shared semantic foundation, while the application profile specifies how that foundation is used in a particular institutional context.

Research on the localization of Schema.org for manuscript description in the Iranian-Islamic information context provides a domain-specific example of this issue. Mohammadi Ostani et al. (2021) examined the vocabulary's ontological structure, including the inheritance of properties from broader types, and proposed additional properties to support manuscript description. Their study illustrates how a general Web vocabulary may require contextual extension to represent the descriptive characteristics of specialized resources. In LLM-based structured data generation, this has a direct implication: the vocabulary and properties selected for an output should be assessed against the conceptual and descriptive requirements of the domain, rather than assumed to be sufficient because they are available in Schema.org.

RAG and knowledge graphs can complement this semantic layer. RAG retrieves information from external sources to support generation, reducing exclusive dependence on model parameters. Knowledge graphs can make entity relationships available for retrieval and comparison. Yet both depend on the quality, currency, and coverage of their sources, and a model can still misinterpret or overstate retrieved evidence (Lewis et al., 2020; Hogan et al., 2021; Pan et al., 2024). For this reason, a system should preserve links between generated claims and the evidence used to support them, assess source authority, and distinguish retrieved facts from inferences or unsupported completions.

The analysis therefore identifies three levels of Schema.org contribution. First, it directly supports semantic description through shared types and properties. Second, it indirectly supports interoperability and error detection by making data elements and relationships more explicit. Third, it does not independently solve generation instability, syntactic validity, factual accuracy, identity resolution, provenance, or application-specific validation. These limits are important for system design: the presence of Schema.org markup

should not be used as a proxy for the trustworthiness of the data it describes.

The distinction between descriptive semantics and operational constraints has implications for metadata design. A vocabulary term can communicate what a property means, but an organization must still decide whether that property is required, repeatable, controlled, or subject to a particular validation rule in a given workflow. Those decisions belong to an application profile and its governance arrangements. A library may require authority-controlled creator identifiers; an archive may need to preserve provenance and contextual relationships; a museum may need object-specific descriptions and links to controlled terminology. The same general vocabulary can be used across these contexts, but the local profile determines how its terms are applied and what additional semantics or constraints are necessary.

A further implication concerns the status of generated data over time. A record that is initially accepted may later become outdated when an organization changes its name, a resource is reidentified, or new evidence corrects a previously accepted value. Trustworthiness therefore requires lifecycle controls: versioning, update policies, change histories, and mechanisms for propagating corrections to linked records. Provenance should record not only the source but also the date and process of acquisition, because a source may be authoritative for one period and obsolete in another. These requirements connect the technical architecture to metadata governance and long-term information stewardship.

A Layered Framework for Trustworthy Structured Data

The proposed framework positions Schema.org within a layered architecture that follows the movement of information from source material to managed and reusable content. The layers are functional rather than strictly sequential: retrieval, validation, and correction may recur as information is enriched or reused. Each layer has a distinct responsibility, and trustworthiness depends on coordination among them.

The source and context layer identifies the documents, databases, authority files, repositories, or other resources from which information may be extracted. It records the context in which a claim is made and, where possible, the authority and temporal status of the source. Source

Articles in Press

selection matters because no downstream vocabulary or validation mechanism can compensate for unreliable evidence.

The processing layer uses an LLM or other extraction mechanism to interpret a request, identify candidate entities, extract values, and generate a provisional description. At this stage, the system should preserve uncertainty and avoid silently filling missing fields with plausible guesses. Candidate entity matches and inferred relationships should remain distinguishable from directly extracted values until they have been checked.

The semantic layer maps candidate entities, attributes, and relationships to shared vocabulary terms. Schema.org can serve as the common descriptive vocabulary at this stage. Where its terms do not provide sufficient specificity, an application profile or a domain vocabulary may be required. The mapping should preserve the distinction between the entity being described, the property assigned to it, and the source or evidence supporting the value.

The representation layer expresses the semantic description in a machine-readable form, such as JSON-LD or RDF. The selected representation must be appropriate for the receiving system and should preserve identifiers and relationships. This layer concerns the serialization and graph representation of data; it should not be confused with the semantic vocabulary itself.

The validation and verification layer performs at least two different forms of control. Structural validation checks whether the representation conforms to syntax, data types, required fields, and application-specific constraints. Factual verification checks whether values and relationships are supported by reliable sources. JSON Schema and SHACL can support constraint validation in their respective representation contexts, while authoritative sources, entity reconciliation, and claim-level evidence checks support factual review. Passing a structural validation test must not be interpreted as proof of truth.

The identity, provenance, and storage layer manages persistent identifiers, authority links, source references, retrieval dates, transformation records, and validation status. Knowledge graphs or other entity-centered repositories can preserve relationships among resources and support retrieval across systems. Provenance should be recorded at an appropriate level of granularity: where possible, each consequential claim or value should be connected to its source and the

process through which it was extracted, inferred, or verified. The W3C PROV ontology provides a general model for representing provenance entities, activities, and agents and can be specialized for application contexts (W3C, 2013).

The retrieval, delivery, and reuse layer makes validated data available to applications, content management systems, repositories, and users. Data may be delivered as structured records, graph statements, or natural-language responses grounded in the stored information. Reuse depends on more than serialization: receiving systems must understand the vocabulary, identifiers, relationships, and application constraints. Feedback from downstream use should be able to trigger correction of earlier processing, mapping, or validation decisions.

Five evaluation dimensions apply across these layers. Syntactic validity asks whether the data can be parsed in the declared format. Structural and semantic conformity asks whether types, properties, value types, and relationships conform to the vocabulary and application profile. Factual accuracy and evidential support ask whether claims are consistent with reliable sources. Provenance and traceability ask whether the origin and processing history of claims can be inspected. Operational reusability asks whether data can be interpreted, exchanged, linked, and used in downstream processes. These dimensions provide evaluation criteria for the architecture; they are not additional technology layers.

The framework treats trustworthiness as a managed quality rather than an intrinsic property of an LLM response. A record may be syntactically valid but semantically inappropriate, semantically well-formed but factually false, or accurate but impossible to audit because its source is unknown. The architecture is intended to make such differences visible and to assign each control to the layer where it can be meaningfully performed.

Articles in Press

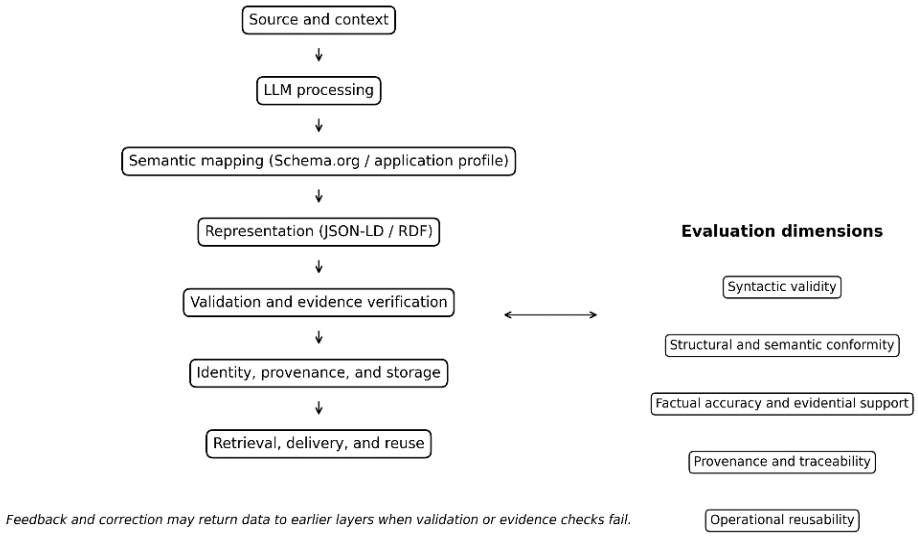


Figure 1. Layered framework for generating and managing trustworthy structured data

Note. The five evaluation dimensions apply across the functional layers; they are not separate technical layers.

Discussion

The findings clarify why the ability to generate structured output should not be equated with the generation of trustworthy structured data. An LLM can produce JSON, tables, or JSON-LD fragments, but the apparent order of the output does not establish that its values are correct, its entities are identified, or its relationships are justified. The distinction is consequential for digital content management because generated descriptions may enter repositories and workflows where they are treated as persistent information rather than provisional conversational text.

Schema.org can improve the semantic organization of generated content by providing shared terms for entities, properties, and relationships. Its value is particularly evident when outputs must be exchanged among systems that recognize the same vocabulary. However, standardization of representation and verification of content

are different operations. Mapping a fabricated name to an author property makes the claim more explicit and processable; it does not make the name correct. The vocabulary can make errors easier to detect, compare, and govern, but cannot independently establish truth.

This distinction also explains why Schema.org should not be treated as a substitute for an ontology or application profile designed to meet every domain requirement. Its broad and flexible model facilitates Web-wide adoption, whereas specialized environments may need stronger constraints, controlled values, authority control, and explicit policies for missing or uncertain information. The appropriate approach is to use Schema.org where its semantics fit, extend or map it where necessary, and document the additional rules through application profiles. This approach avoids both underestimating the vocabulary's interoperability value and attributing to it functions it was not designed to perform.

The relationship between general-purpose Web vocabularies and domain-specific content standards is particularly relevant to bibliographic data. Rouhi et al. (2020), in their study of interoperability between Schema.org and Resource Description and Access (RDA), examined how RDA's content guidelines and instructions for value assignment can complement the descriptive properties of Schema.org. Their analysis highlights that mapping a bibliographic element to a Schema.org property does not, by itself, determine how its value should be interpreted or assigned in a particular cataloguing context. This supports the use of domain-specific content standards alongside general Web vocabularies when generating structured data for specialized information environments.

The framework also highlights the importance of entity-centered information management. When records are treated as isolated bundles of fields, identity and relationship errors can remain difficult to detect. Explicit entities, persistent identifiers, authority links, and graph relationships provide a basis for reconciling descriptions across sources. Yet identity links must be evidence-based. The use of sameAs or another relationship property without adequate reconciliation can create false equivalences and contaminate connected datasets.

For organizations, trustworthiness requires policies that govern not only the choice of vocabulary but also source selection, authority control, validation, provenance, and correction. A metadata pipeline should specify who or what may assert a value, what evidence is

Articles in Press

required, how uncertainty is represented, which constraints are mandatory, and how corrections are propagated. This is particularly important in libraries, archives, museums, research repositories, and public information systems, where descriptions may be retained and reused over long periods. Data governance and metadata governance therefore become practical conditions for the responsible use of generated structured data.

The proposed framework is analytical and conceptual. It synthesizes documentary and technical knowledge to clarify functions and boundaries; it does not demonstrate that a particular combination of technologies will improve accuracy by a measured amount. Its propositions should be tested empirically across models, domains, and data types. Such evaluation should report separate measures for syntax, semantic mapping, identity resolution, factual correctness, provenance completeness, and downstream reusability rather than collapsing them into a single undifferentiated trust score.

Conclusions and Implications

This study analyzed challenges in generating trustworthy structured data with LLMs and examined the role and limitations of Schema.org. The analysis organized 18 challenges into five categories: generation and stability; structure and conformance; semantics and identity; accuracy and evidence; and provenance and traceability. It showed that Schema.org contributes most directly to shared semantic description through common entity types, properties, and relationships. This contribution can improve consistency, interoperability, and the potential reuse of generated data, but it does not independently guarantee syntactic validity, factual accuracy, correct identity, relationship validity, or provenance.

The proposed layered framework assigns complementary functions to source retrieval, language-model processing, semantic mapping, machine-readable representation, validation and verification, identity and provenance management, and downstream retrieval and reuse. It evaluates the resulting data through five dimensions: syntactic validity; structural and semantic conformity; factual accuracy and evidential support; provenance and traceability; and operational reusability. The central implication is that trustworthiness must be assessed across the lifecycle of data creation and use. A standardized description is an

important resource for managing meaning, but it cannot substitute for evidence, authority control, validation, or organizational governance.

For practice, organizations should define application profiles and data-quality rules for their domains, use persistent identifiers and authoritative sources for entity resolution, validate representations with appropriate constraint languages, and preserve claim-level provenance wherever feasible. LLM-generated values should remain provisional until they have passed the relevant controls. When evidence is absent, systems should preserve unknown or unverified status rather than convert uncertainty into an apparently complete record.

Future research can compare LLMs in generating Schema.org-based descriptions, evaluate how retrieval and knowledge graphs affect different error categories, and test the framework in library, archival, museum, publishing, and government contexts. Such studies should distinguish improvements in representational consistency from improvements in factual accuracy and should assess the cost and effectiveness of human review and automated controls.

Articles in Press

References

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S., & Zagni, G. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. Nature Machine Intelligence, 6(8), 852–863. https://doi.org/10.1038/s42256-024-00881-z

Berners-Lee, T. (2006). Linked data. https://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22. https://doi.org/10.4018/jsws.2009081901

Cyganiak, R., Wood, D., & Lanthaler, M. (Eds.). (2014). RDF 1.1 concepts and abstract syntax. World Wide Web Consortium. https://www.w3.org/TR/rdf11-concepts/

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Morgan & Claypool. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4), Article 71. https://doi.org/10.1145/3447772

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv. https://doi.org/10.48550/arXiv.2311.05232

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

JSON Schema. (2020). JSON Schema Core and Validation specifications (Draft 2020-12). https://json-schema.org/specification

Lavrinovics, E., Biswas, R., Bjerva, J., & Hose, K. (2025). Knowledge graphs, large language models, and hallucinations: An NLP perspective. Journal of Web Semantics, 85, Article 100844. https://doi.org/10.1016/j.websem.2024.100844

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://arxiv.org/abs/2005.11401

Mohammadi Ostani, M., Cheshmeh Sohrabi, M., Taheri, S. M., Shabani, A., & Asemi, A. (2021). Localization of Schema.org for manuscript description in the Iranian-Islamic information context. Knowledge Organization, 48(5), 345–356. https://doi.org/10.5771/0943-7444-2021-5-345

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

Rouhi, E., Taheri, S. M., Alipour-Hafezi, M., & Samiei, M. (2026). Interoperability between Schema.org and RDA for describing bibliographic entities. Knowledge Organization, 53(1), Article 43984. https://doi.org/10.31083/KO43984

Schema.org. (n.d.-a). Data model. https://schema.org/docs/datamodel.html

Schema.org. (n.d.-b). How we work. https://schema.org/docs/howwework.html

Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering, 27(2), 443–460. https://doi.org/10.1109/TKDE.2014.2327028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762

W3C. (2013). PROV-O: The PROV ontology. https://www.w3.org/TR/prov-o/

W3C. (2017). Shapes Constraint Language (SHACL). https://www.w3.org/TR/shacl/

W3C. (2020). JSON-LD 1.1: A JSON-based serialization for linked data. https://www.w3.org/TR/json-ld/

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., & Li, B. (2024). DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. arXiv. https://arxiv.org/abs/2306.11698

Articles in Press