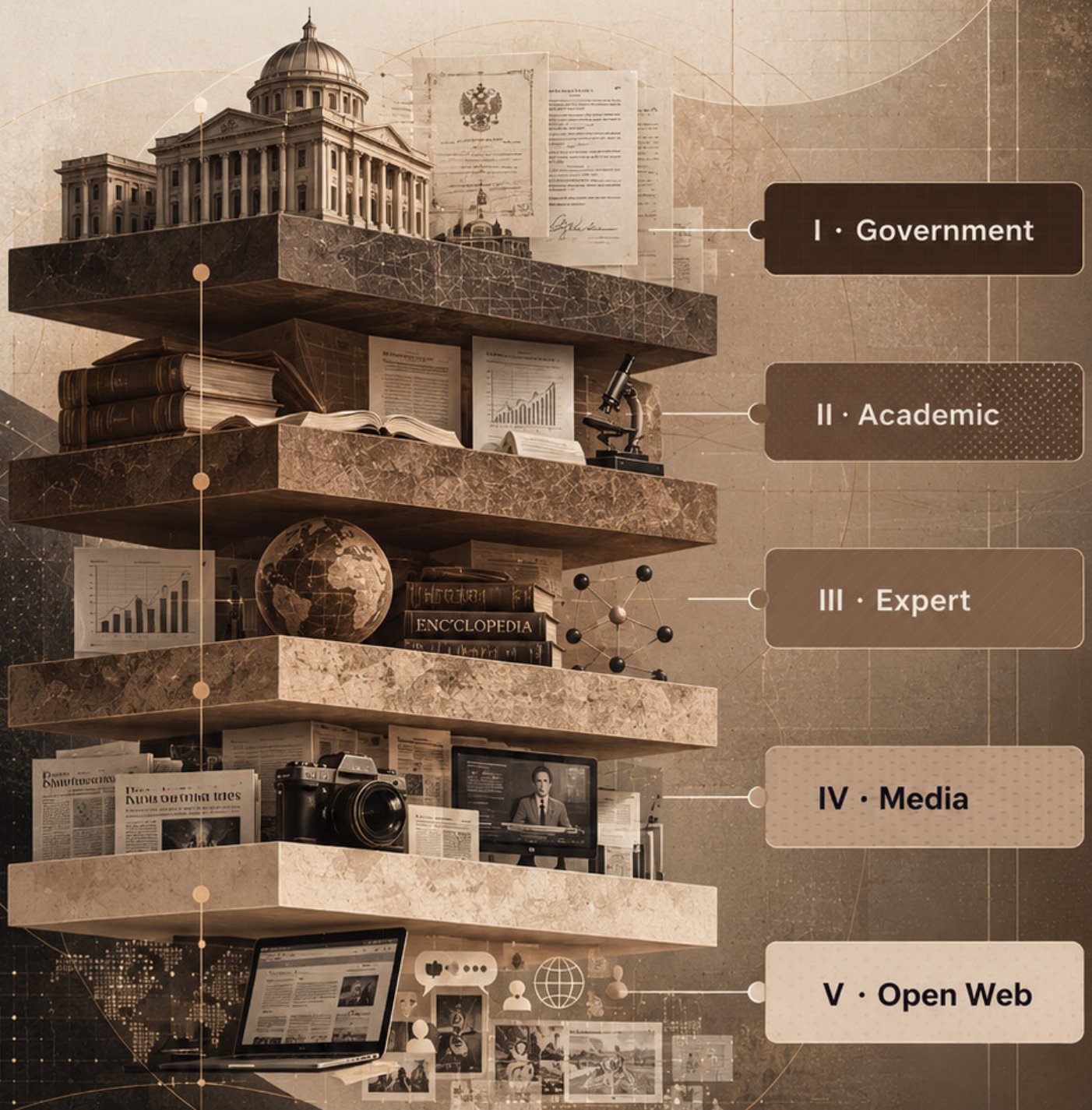


Information Stratigraphy

A FIVE-LEVEL MODEL OF SOURCES
BEHIND AI VISIBILITY



Information Stratigraphy: A Five-Level Model of Sources Behind AI Visibility

Serhii Pontus

QuorumLab, Finland

✉ Email: sp@quorumlab.eu

Leonid Kachuliak

QuorumLab, Finland

✉ Email: leonid.kachuliak@quorumlab.eu

Abstract

The spread of generative artificial intelligence is changing how information about organisations, companies and individuals is formed and circulated. Instead of a list of documents to examine for themselves, users increasingly receive a synthesised answer that draws on many sources at once. Research on AI Visibility has mostly examined the observable output of this exchange: mentions, citations, recommendations and the properties of answers. Yet the public information environment that surrounds a given entity before anyone queries an AI system can itself be treated as an object of scholarly study.

This article proposes the QuorumLab Information Stratigraphy Model (QISM), a five-level heuristic model of an entity's public information environment. It distinguishes government, academic, expert, media and open-web levels, and its main classificatory principle is the institutional procedure to which a document is subject. The model separates the document, the claim it contains and the producer of the information. It also takes account of provenance, affiliation, currency and the technical conditions under which AI systems gain access to sources.

The theoretical basis draws on research into gatekeeping, cognitive authority, the social organisation of information flows and the sources of generative search. We also review studies showing that training data, document characteristics and source identity matter for generative answers.

QISM has both scholarly and practical value. It allows a researcher to build a structured information map of an organisation or person, trace where information comes from, identify discrepancies and formulate new research hypotheses. The model sets out a sequential research programme that moves from describing the public environment to comparing it with AI Visibility, examining potential influencing factors and testing them empirically.

The paper is conceptual and methodological. It presents a heuristic model intended as a foundation for subsequent research.

Keywords

AI Visibility, information stratigraphy, QISM, information environment, gatekeeping, cognitive authority, public sources, generative artificial intelligence, information provenance.

Introduction

Generative artificial intelligence is changing the way people obtain information about the world. A conventional search engine gives the user a list of sources to read and compare. A generative system can present the same information as a single answer that combines facts, interpretations and recommendations.

When a user asks about a company, an organisation or a particular person, the answer may include information about the entity's activities, leadership, professional reputation, achievements, legal status or public standing.

The underlying information environment, however, can be highly heterogeneous. Information about a single entity may appear simultaneously in government registers, academic publications, analytical reports, journalism, encyclopaedias, corporate documents and user discussions.

Documents that contain the same claim do not necessarily share the same origin. They may have passed through different verification procedures, rely on one another, reproduce outdated data or refer to different legal and factual circumstances.

A generative answer is therefore an observable result behind which a complex public information environment may lie.

Research on AI Visibility has developed mainly around measuring this result. Studies examine how often companies are mentioned, how they feature in recommendations, how sources are cited, how the visibility of documents changes and how answers differ across models.

This work is necessary, but it leaves a separate question open: how is the public information environment of a given entity itself structured, regardless of which of its elements an AI system has used? The question matters in its own right.

An organisation may be documented in detail in its own corporate materials yet be almost absent from independent professional publications. Another may have many press mentions, all of which go back to a single press release. A third may have contradictory information in official registers and in derivative sources.

In each case, a simple count of documents tells us little about how the environment is structured. Studying that structure requires an approach that can establish the institutional origin of

documents, the links between them, the degree of their independence and the ways in which claims spread.

This article proposes the QuorumLab Information Stratigraphy Model (QISM), a heuristic model for classifying an entity's public information environment.

The name draws on a geological analogy. Stratigraphy helps geologists describe how rock layers are arranged; information stratigraphy treats the public information about an entity as a structured system of distinct institutional levels.

The analogy does not imply that geological regularities carry over literally to information. Its purpose is to give the researcher a scheme for describing a complex object systematically.

Contribution of the article

The article makes four interrelated contributions.

The first is a five-level model of the public information environment of a specific entity, distinguishing government, academic, expert, media and open-web levels.

The second is a principle for classifying a document by the institutional procedure to which it is subject, grounded theoretically in gatekeeping research.

The third is a separation between the document, the claim and the producer of information. This makes it possible to study how information originates and changes as it moves between levels.

The fourth is a separate dimension, the AI Access Pathway, which describes the technical and contractual routes by which AI developers obtain public information.

Together these form a single heuristic framework. It can be used to build an information map of an entity and as a basis for later research into why generative answers take the form they do.

Methodology and theoretical foundations

1. Methodology of the conceptual study

This is a conceptual study: its outcome is a theoretical model rather than an empirical assessment of an information environment. The model was built in four consecutive steps.

The first step was selecting the theoretical foundations. We examined work that describes the information environment as heterogeneous and institutionally organised (gatekeeping theory, cascade and layered models of information flow, and the theory of cognitive authority), together with research on the sources of generative answers. A work was included if it bore directly on one of three questions: how the institutional conditions of information production differ, how the authority of a source is established, and which sources generative systems draw on. The review is targeted rather than systematic.

The second step was adapting the five-level structure. The five-level pyramid of public opinion formation (Maltsev, 2023) served as a heuristic framework. Its levels were redefined as classes of public documents about an entity rather than as levels of opinion.

The third step was formulating the classification rule. Drawing on gatekeeping theory, we set out a single principle for assigning a document to a level according to the institutional filter to which

it is subject. The rule was examined analytically against boundary cases (sponsored content, preprints, openly edited encyclopaedias, structured knowledge bases and AI-generated content) and refined to account for different types of filter.

The fourth step was introducing cross-cutting dimensions. Properties of a document that cannot be expressed consistently through its level (affiliation, platform, provenance, filter type, mode of production and AI access) were treated as separate characteristics.

Section 5 applies the model to a hypothetical example. Empirical testing belongs to the research programme set out in Section 7.

2. The social organisation of information flows

An information environment is not a collection of equivalent messages. Information is produced, selected and disseminated under different institutional conditions.

Government bodies, research institutions, professional communities, newsrooms and users of digital platforms each work within their own procedures, rules and forms of accountability.

Gatekeeping theory is one of the theoretical traditions that can be used to study these differences.

David Manning White (1950) studied how a newspaper editor selected news and showed that incoming information passes through decisions to include or exclude material. Shoemaker and Vos (2009) later synthesised gatekeeping research and examined how individual, organisational and institutional factors shape information flows.

For this article, the existence of selection matters less than the differences between the procedures that carry it out.

Peer review, judicial proceedings, professional certification and journalistic editing are different mechanisms. Each of them, however, sets conditions under which information takes on a public form and acquires a corresponding institutional status. QISM uses this idea to classify documents.

On this reading, the levels of the information environment correspond to different types of institutional gatekeeper. Applying the concept of gatekeeping beyond its original media context is a conceptual extension of the theory.

3. Earlier models of information levels

In his model of *cascading activation*, Robert Entman (2003) described how political interpretations spread between institutional actors in public communication. The model shows that information processes can be organised on several levels and that the content of messages changes as it passes from one actor to another.

Gunther Eysenbach (2020) proposed the *Information Cake* model, which distinguishes four layers of the information environment: science; policy and practice; news media; and social media. The publication is a conceptual piece that the publisher labels as non-peer-reviewed.

The model was developed in the context of infodemiology. Its value lies in showing how information layers differ in their procedures for checking content and in the ways knowledge spreads.

Among the antecedents of QISM, a special place belongs to the five-level pyramid of public

opinion formation attributed to Massimo Introvigne and set out by Oleg Maltsev (2023). That account distinguishes five levels: the opinion of the state, academic opinion, expert opinion, media opinion and the opinion of social media users.

The pyramid serves as the heuristic starting point for QISM. There is, however, a fundamental difference between the original idea and the model proposed here.

The public opinion pyramid describes the levels at which social perceptions are formed. QISM uses the five-level structure to classify specific public documents about an entity. It adds a rule for determining a document's level, a system of sublayers, separate document characteristics and provenance relations between claims.

QISM therefore does not transfer an existing scheme mechanically. It builds a separate instrument for studying the public information environment.

We were unable to locate, in the literature reviewed, the original work by Introvigne that would establish authorship and the first formulation of the five-level pyramid. This article therefore cites the account given by Maltsev (2023).

4. Cognitive authority and institutional status

The second relevant line of theory concerns cognitive authority.

Wilson (1983) examined the conditions under which information received from other people and institutions comes to be accepted as trustworthy. Rieh (2002) studied how people judge information quality and cognitive authority on the web. Hilligoss and Rieh (2008) proposed a framework for credibility assessment that takes the context of interaction into account. Hirvonen et al. (2025) reviewed empirical research on cognitive authority, including its social, institutional and communicative aspects.

This body of work supports an important distinction: the institutional status of a document is not the same as the reliability of the claims it contains.

A government decision may be overturned. A peer-reviewed article may contain an error. A news report may accurately describe an event that has not yet appeared in official records. A user's post may contain reliable testimony.

The order of the information levels cannot, therefore, be read automatically as a scale of truth. QISM treats a document's level primarily as a description of the institutional setting in which it was produced and checked.

5. Sources of generative answers

Research on generative search provides an empirical basis for studying the heterogeneity of information sources.

Li and Sinnamon (2024) analysed the sources cited by generative search engines and found that different categories of source were represented unevenly.

Nguyen et al. (2026) examined the sources generative systems use when answering questions about mental health. The authors analysed 15,942 citations from 1,713 unique domains using a nine-category organisational typology.

These studies show that sources can be systematised, although they mostly begin their analysis

from generative answers that have already been produced.

Studies of company information add a further perspective. Analysing citations in generative systems, Žatuchin (2026) found that in one component of the dataset 77.0% of 131,514 URL citations fell into the residual category *other_web*. A concentration this high shows that the original typology had limited power to discriminate between sources.

Kumar (2026) analysed 149,912 citations in answers mentioning the brands under study and found that 75.2% went to the corporate pages of third parties, including competitors, suppliers and other commercial organisations.

These findings relate to particular samples and cannot be generalised automatically to all AI systems and query types. They do show, however, that categories of public source need closer analysis.

The distinction between classifying a domain and classifying a document is especially important. A single website may host independent editorial content, corporate press releases, user posts and automated material at the same time. Knowing the domain, then, does not always reveal where a particular claim actually came from.

6. How source characteristics affect generative answers

Several experimental studies confirm that the underlying information environment matters.

Kandpal et al. (2023) found a relationship between the number of relevant documents in a language model's training data and the accuracy of its answers to factual questions. The study includes a causal analysis and shows that how well information is represented in the training corpus affects what the model learns.

In *GEO: Generative Engine Optimization*, Aggarwal et al. (2024) showed that changing the content of web documents can affect how prominently they appear in generative answers. Under experimental conditions the authors recorded relative improvements in visibility of up to 40% on their research metrics, including Position-Adjusted Word Count. This result concerns changes to the content of documents already available to a generative search engine in a given context. It does not mean a 40% higher chance of being discovered or cited in commercial AI products.

In controlled experiments with political news, Dai et al. (2025) showed that the choice of which source to cite can depend on the identity of the media outlet and not only on the content of the article.

This result is of particular interest for QISM. It means that the same content, placed in a different context of origin, will not necessarily be treated in the same way by a generative system.

The distinction between a document and a claim therefore has research value of its own. It should be noted, though, that Dai et al. studied political news and differences between media sources, not the full five-level structure proposed here.

Hamann et al. (2026) also examined how language models update their knowledge of brands and how the conditions of information retrieval affect the answers they produce.

Taken together, these studies show that the content, representation and provenance of sources are significant variables in AI Visibility research. QISM brings these lines of work together in a conceptual model of an entity's underlying public information environment.

7. Comparison with earlier approaches

Table 1. Conceptual comparison of QISM with earlier models and studies

Approach	Primary unit of analysis	Basis of division	Layering or hierarchy	Main object of study	Provenance of claims
Entman (2003)	Actors and communication processes	Political influence and the spread of interpretations	Cascade model	Political communication	Through the spread of interpretations
Eysenbach (2020)	Information environments and processes	Institutional role of information producers	Four layers	Infodemics and the movement of knowledge	Knowledge translation between layers is considered
Five-level pyramid attributed to Introvigne (as reported by Maltsev, 2023)	Levels of public opinion	Social and institutional affiliation	Five levels	Formation of public opinion	Not formalised at document level
Li and Sinnamon (2024)	Cited sources and domains	Source type and institutional affiliation	Typology of categories	Sources of generative answers	Not a primary unit of analysis
Nguyen et al. (2026)	Cited domains	Organisational category	Nine categories	Generative answers on mental health	Not a primary unit of analysis
Seo et al. (2026)	Citations and their correspondence with answers	How a source supports and is used in the answer	Analytical classification	Citation quality of AI systems	Link between answer and source is considered
QISM	Document, claim, producer	Institutional filter to which the document is subject	Five levels with sublayers	An entity's public environment, prior to analysis of AI answers	Separate provenance and versioning links

The comparison shows that QISM does not claim priority in distinguishing information levels or classifying sources.

What sets it apart is the way it combines entity-centred analysis of the public environment, classification of documents by institutional procedure, and treatment of claims as distinct units of information that move between documents.

The QuorumLab Information Stratigraphy Model

1. Scope and definition

The QuorumLab Information Stratigraphy Model (QISM) is a heuristic five-level model for describing the public information environment of a specific entity in a structured way.

The entity may be a company, an organisation, a person or any other identifiable subject about which public information exists. The information environment is understood as the set of documents, claims and information producers, together with the links between them.

QISM studies this environment as it exists independently of how any particular AI system processes the information in it. The distinction is essential: the model describes the potential

informational basis from which answers are formed, not the internal algorithms of AI systems.

2. *Three core entities*

The model distinguishes three types of object of analysis.

A *document* is a published information object that contains information about the entity under study and can be identified as a discrete unit.

A *claim* is a specific substantive statement about the entity, such as a founding date, the name of a chief executive, the existence of a licence, a description of activities or any other verifiable statement.

An *information producer* is the person, organisation or institution responsible for creating, preparing or publishing a document.

Keeping these apart avoids a common analytical conflation. One document may contain many claims. One claim may be reproduced in several documents. One producer may publish documents subject to different verification procedures.

A researcher, for instance, may publish a peer-reviewed article, an opinion column in a newspaper and a personal post on social media. Their professional qualifications do not make these documents equivalent in institutional origin.

The primary QISM classification therefore applies to the document, while the characteristics of the producer are recorded separately.

3. *The institutional filter principle*

The core rule of the model is as follows:

A document is assigned to the QISM level corresponding to the qualifying institutional procedure governing its creation, publication or subsequent content control.

An institutional filter is a procedure of selection, registration, expert review, verification or editorial approval that determines the conditions under which a document exists publicly.

The model distinguishes three types of institutional filter. *Pre-publication review* checks content before it appears, as in peer review or editorial selection. *Institutional certification of status* gives a document legal or official force, as with state registration, a court ruling or a regulatory decision. *Post-publication control* means that published material remains subject to review under codified rules and can be corrected or removed, as in openly edited encyclopaedias and structured knowledge bases.

A document's level is determined by the qualifying institutional procedure to which it is subject, rather than by the status of its producer or hosting platform. The type of institutional filter is recorded as a separate characteristic. Wikipedia, where checking takes place after publication, and a peer-reviewed article, checked beforehand, are therefore covered by the same rule without exceptions.

The status of the producer, the status of the hosting platform and the status of the document itself are kept apart. An independent journalist's article, for example, may pass editorial review and

belong to the media level. A paid press release on the same website that has not been through independent editorial selection, however, remains affiliated material.

The reputation of a platform should therefore not be transferred automatically to every document it hosts.

For the purposes of QISM, a codified system for assessing the content of information must be distinguished from the general moderation of a digital platform. The former involves publicly established rules for including, verifying, changing or removing information, a documented history of editorial actions, and procedures for discussing or reviewing decisions. Applying such rules requires specialist knowledge and familiarity with the relevant institutional practice.

General moderation may chiefly regulate the behaviour of participants, what may be posted and compliance with community rules. Its decisions need not come with a public justification, and the history of how they were reached may not be available to an outside researcher.

Having rules and the ability to remove material is not, in itself, sufficient grounds for placing a platform on the expert level of QISM.

4. The five levels of the information environment

QISM distinguishes five main levels:

- Government
- Academic
- Expert
- Media
- Open Web

Their arrangement follows the original five-level pyramid of public opinion.

The order reflects a proposed heuristic sequence of institutionalisation and public accountability, running from formally regulated government procedures, through scholarly and professional mechanisms, to editorial review and open self-publication.

This sequence is not a measured, universal scale of rigour or truth. Peer review and legal registration assess different properties of documents. Expert analysis may go deeper than a piece of journalism, but that does not make every expert document more reliable than every media publication.

The QISM hierarchy therefore describes the presumed institutional structure of the public information environment. It does not assert that some sources are universally superior to others.

The five levels are intended to provide formal coverage, with Open Web serving as the residual category for documents that do not meet the qualifying procedural criteria of the four higher levels. Whether the sublayers are complete in substance is a separate research question.

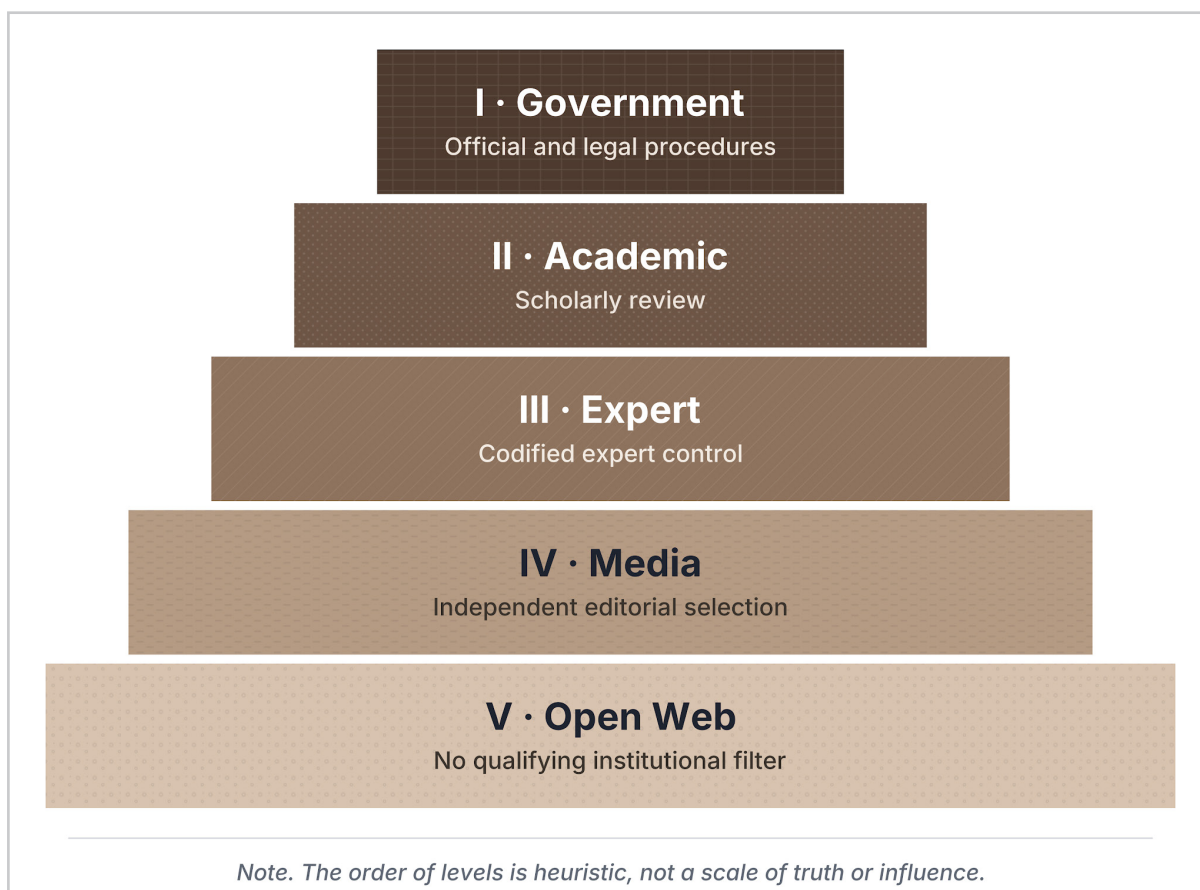


Figure 1. The five-level QISM stratigraphy

5. *The government level*

The government level brings together documents created or approved under official powers, legal procedures and public legal accountability. We propose five sublayers:

- Legislation and court decisions.
- Government registers.
- Regulatory decisions and official administrative documents.
- Official statistics.
- Public statements by officials and government press offices.

The sublayers differ in the legal nature of the document. A registration entry, for example, confirms a particular legal fact but does not always describe the organisation's earlier history. A statement by an official, for its part, does not carry the status of a legal act.

The order of the sublayers is part of the heuristic construction and may be refined in later research.

6. *The academic level*

The academic level includes documents that have passed an established scholarly editorial or examination procedure. Three sublayers are proposed:

- Scholarly publications with external peer review.

- Work assessed by an academic committee, including defended dissertations.
- Scholarly publications with editorial review but no external reviewers.

The type of review (double-blind, single-blind or open) is recorded as an additional characteristic of the procedure. Journal quartile, indexing, publication format and publisher status are also recorded separately.

This avoids equating a journal's bibliometric standing with the rigour of the review applied to a particular paper.

The main criterion for the academic level is the presence of an established scholarly procedure. A publisher's assertion that a publication is scholarly is not enough.

7. The expert level

The expert level brings together documents associated with professional assessment, specialist competence or a codified community verification procedure. Seven sublayers are proposed:

- Standards and accreditation documents.
- Audit opinions, professional ratings and certification.
- Encyclopaedias with codified verifiability requirements.
- Structured knowledge bases with rules for including entities.
- Analytical work by research and industry organisations.
- Expert registers, awards and rankings with established selection.
- Signed expert opinions and moderated preprints.

Wikipedia and Wikidata occupy a distinctive position.

English Wikipedia has a system of rules on verifiability and the use of reliable sources (Wikipedia contributors, n.d.-b). The *Reliable Sources* page and its associated guidelines set out how sources may be used in encyclopaedic text.

The *Perennial Sources* information page summarises recurring discussions of particular sources, but it is not itself a Wikipedia policy (Wikipedia contributors, n.d.-a). It documents a practice of collective source assessment that takes the context of use into account.

Wikidata applies adopted notability rules for entities and other mechanisms of community review (Wikidata contributors, n.d.-a). Its *Verifiability* page, which states that claims require sources, has the status of a *proposed guideline* rather than adopted policy (Wikidata contributors, n.d.-b).

The verification procedures of Wikipedia and Wikidata cannot, therefore, be treated as identical. In QISM, Wikipedia belongs to the sublayer of encyclopaedias with a formalised verification procedure, and Wikidata to a separate sublayer of structured knowledge bases.

The difference reflects the nature and scope of the rules in force. It does not presume that one platform is more reliable than the other. It also allows the institutional structure of open communities to be examined without automatically equating community moderation with peer review.

Wikipedia and Wikidata represent different forms of codified information governance. Wikipedia has well-developed rules on the use of sources, a public edit history and procedures for discussing content. Wikidata has its own rules for including entities, a history of changes

to structured data and mechanisms of community oversight; its requirements for sourcing individual claims differ from Wikipedia's.

Reddit also has community rules, a posting history and appeal mechanisms. Much of its moderation, however, is carried out within individual communities, does not necessarily involve checking the substance of claims and is not always accompanied by a public record of the reasons.

The line between the expert and open-web levels is therefore drawn by the nature, object and documentability of institutional control over information, rather than by the presence of moderation or the complexity of an interface. What differs is the system of information governance. This does not mean that every claim on Wikipedia or Wikidata has in fact been verified, or that every moderation decision on Reddit is arbitrary.

8. The media level

The media level includes documents that have passed editorial selection by professional publishers but are not academic publications. Five sublayers are proposed:

- International news agencies and leading publications.
- National media and broadcasters.
- Trade and specialist media.
- Regional and local press.
- Non-academic book publishers with editorial selection.

Whether a document belongs to the media level depends on the editorial process it actually went through. This matters particularly for commercial and sponsored content.

Material placed in the press-release section of a well-known publication, where the client publishes it directly, does not necessarily go through the same editorial procedure as an independent news article.

QISM thus makes it possible to distinguish the institutional status of a platform from the procedure by which a specific item actually came into being. The distinction is relevant to sociological research on information transparency and to the analysis of disguised advertising, corporate rewrites and other forms of information influence.

9. The open-web level

The fifth level covers public material that does not meet the qualifying institutional-procedure criteria of the four higher levels. It includes four sublayers:

- Platforms with community rules and limited moderation.
- Self-publication by organisations and individuals.
- User-generated content.
- Aggregators, automated directories and copies of material.

This level may include corporate websites, official accounts, technical documentation, forums, personal blogs, reviews and other material.

Self-publication does not imply unreliability. A company's official website, for example, may be

the primary source for its own contact details, product specifications or public statements.

Information posted by an organisation itself nevertheless differs in origin from an independent professional publication about it, and research on information environments needs to take this into account.

I Government	II Academic	III Expert	IV Media	V Open Web
Legislation and court decisions Government registers Regulatory decisions Official statistics Statements by officials	Peer-reviewed publications Committee-assessed theses and dissertations Editorially reviewed work	Standards and accreditation Audits, ratings, certification Encyclopaedias Structured knowledge bases Research and industry analysis Registers, awards and rankings Signed expert opinions and moderated preprints	International agencies National media and broadcasters Trade and specialist media Regional and local press Non-academic publishers	Community platforms Self-publication User-generated content Aggregators and copies

Note. Sublayers and their order are heuristic and open to empirical refinement.

Figure 2. Sublayers of the five institutional levels in QISM.

10. Additional document characteristics

The five-level classification assigns each document to an institutional level but does not capture all of its properties. QISM therefore uses additional cross-cutting dimensions.

Table 2. Additional QISM characteristics

Characteristic	Content
Affiliation	Relationship between the producer and the entity
Platform	Type and institutional status of the hosting resource
Provenance	Primary document, paraphrase or derivative material
Currency	Date of publication, update, obsolescence or withdrawal
Versioning	Relationship to other versions of the document
Language and format	Language; text, audio, video or structured data
Procedure confirmation status	Confirmed, presumed or unconfirmed
Filter type	Pre-publication review, certification of status or post-publication control
Academic status	Peer review, publication type, indexing
Mode of production	Human, AI-generated or mixed
AI Access Pathway	Availability of the platform to a specific AI developer

AI-generated content needs a separate note. In QISM it does not form a level of its own, because the way a text is created and the institutional procedure by which it is published are different characteristics.

AI-generated material may be posted without any external check, pass through journalistic editing or be used within some other documented procedure. The mode of production is therefore recorded as a separate label.

11. The infrastructure of AI access to sources

Apart from the institutional status of a document, there is a separate question: how can an AI developer obtain the information it contains? To describe this, we propose the *AI Access Pathway* parameter.

The parameter describes the documented technical and contractual routes by which a specific AI developer can access material from a given platform. These routes include ordinary public access, retrieval by search mechanisms, specialised APIs and contractual data-supply channels.

The values are not mutually exclusive. A publicly accessible platform may also supply material through a commercial API.

In 2024, for example, Reddit announced an expanded partnership with Google that provides access to the Reddit Data API. Google's official announcement stated that this access would allow Reddit data to be displayed, used and applied more effectively, including for model training (Reddit, 2024a; Google, 2024).

The same year, Reddit and OpenAI announced a partnership giving programmatic access to current Reddit content for use in OpenAI products (Reddit, 2024b).

Another example is Wikimedia Enterprise, which offers specialised tools for accessing the freely licensed content of Wikimedia projects. In January 2026 Wikimedia announced new partners for the service: Amazon, Meta, Microsoft, Mistral AI and Perplexity (Wikimedia Enterprise, 2026a). Google has used Wikimedia Enterprise since 2022. The Wikimedia Enterprise documentation also provides for access to Wikidata data (Wikimedia Enterprise, 2026b).

Licensing content and providing a technical access service are different matters. Wikipedia text is distributed under free licences, including CC BY-SA, whereas Wikidata's structured data is published under CC0.

The existence of such a service does not mean that a particular document has been used by an AI system in training or in generating an answer.

Four circumstances must therefore be kept apart: the existence of a formal agreement; the technical possibility of obtaining data; the actual use of a specific document in training or retrieval; and the influence of that document on a generative answer. These are not equivalent, and establishing one of them does not prove the others.

The AI Access Pathway is an additional dimension of the information environment, not a sixth level of QISM. Its value lies in allowing researchers to examine potential differences between documents that sit at the same institutional level but differ in how technically accessible they are.

	Origin Affiliation Provenance Mode of production	Time Currency Versioning	Form Language and format	Procedure Procedure confirmation status Filter type Academic status	Platform Platform type AI Access Pathway
I · Government	●	●	●	●	●
II · Academic	●	●	●	●	●
III · Expert	●	●	●	●	●
IV · Media	●	●	●	●	●
V · Open Web	●	●	●	●	●

Note. Attributes apply where applicable at any level; they are not a sixth level.

Figure 3. Cross-cutting characteristics of documents across QISM levels.

12. The provenance of claims

QISM treats claims as units of information that can move between documents at different levels. These links can be described with the W3C PROV-DM conceptual model (W3C, 2013), which includes relations of provenance, derivation, usage and attribution.

Information published by an organisation, for instance, may be taken up by a journalist, then reproduced in an encyclopaedia and later find its way into other information resources. As it spreads, a claim may stay the same, be made more precise or change.

The presence of the same claim in two documents does not by itself prove that one borrowed it directly from the other. Establishing such a link requires citations, an edit history or some other documented basis.

Information stratigraphy thus allows a researcher to study both how documents are distributed across levels and the possible routes by which the information in them spreads.

13. A short procedure for determining the level

The proposed heuristic procedure can be set out as a sequence of questions.

Question 1. Has the document passed an established government or official legal procedure? If so, it belongs to the government level.

Question 2. Has the document passed a confirmed scholarly editorial or examination procedure? If so, it belongs to the academic level.

Question 3. Is the document subject to specialist professional assessment, or to a codified system of content control with public rules for handling claims, a history of changes and mechanisms for correcting or reviewing them? If so, the document is considered within the expert level. Behavioural moderation alone is not sufficient.

Question 4. Has the document passed independent editorial selection by a professional media outlet or publisher? If so, it belongs to the media level.

If none of these procedures can be established, the document belongs to the open-web level. Where the case is ambiguous, the uncertainty about the procedure should be recorded, and the status of the producer and of the platform noted separately.

This is an outline of how the model is applied. A full procedure for resolving boundary cases is left for a future codebook.

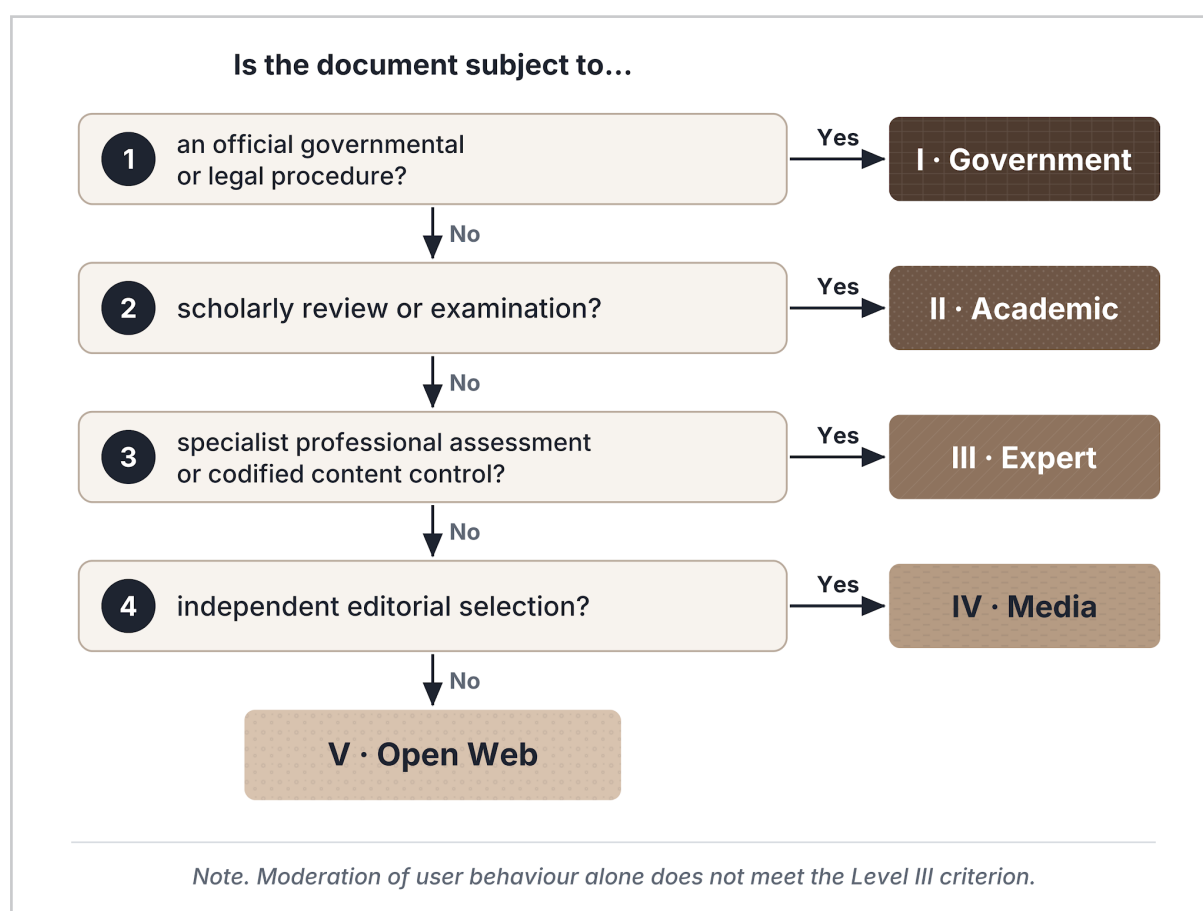


Figure 4. The QISM decision procedure for assigning documents to institutional levels.

Scholarly and practical significance of QISM

1. The information map of an entity

The main value of QISM is that it allows the public information environment of a specific entity to be presented as a structured system. A company, an organisation or a researcher can look at the whole body of available sources, distributed across institutional levels, rather than at

individual publications.

The practical output of such an analysis can be an *information map* of the entity. The map shows which documents exist about the entity, how they are distributed across levels, who produces the information, and which information is primary and which is reproduced through derivative sources.

A researcher may find, for instance, that information about an organisation is widely available on its own corporate resources but almost absent from independent professional publications. In another case, a large number of news items may rest on a single primary document. In a third, the information in a government register may differ from what media outlets have repeated many times.

QISM makes it possible to identify such differences and to ask where they came from.

The absence of documents at a particular level is not treated automatically as a deficiency. What it means depends on the nature of the entity, its activities and the research task at hand.

2. From observing AI Visibility to studying the information environment

Measuring AI Visibility shows how an entity is represented in the answers of generative systems. QISM adds a separate object of observation: the public information environment surrounding the entity.

Two distinct descriptions thus become available. The first is what AI systems report about the company, organisation or person. The second is which public documents and claims about that entity exist independently of AI systems.

Comparing the two opens up new questions. Why do some facts appear in generative answers while others do not? Why is an outdated claim reproduced more often than a current one? What role do the primary document and later paraphrases play? Does the use of material differ according to the level of the source, its language, the editorial procedure or the mode of technical access?

QISM does not answer these questions. Its scholarly value is that it allows them to be posed more precisely.

3. Value for organisations, researchers and independent specialists

The model can be used by different participants in the information environment. For researchers, it provides a basis for describing public sources systematically and formulating testable hypotheses. For organisations, it offers a structured view of the public information that exists about them, where it comes from and how it is connected. For independent specialists, it gives a consistent scheme for deciding which documents to examine when analysing an entity's information presence.

Importantly, applying QISM does not require prior access to the closed algorithms of AI systems. The model treats the public information environment as an independent object of analysis.

Identifying contradictions, outdated information or structural gaps does not, however, mean that intervening in the relevant documents will automatically change generative answers. That step requires separate research and measurement.

A hypothetical example

Consider, for illustration, a hypothetical Company X.

Suppose that its official website states that it began operating in 2015. A journalist at a trade publication uses the company's information and gives the same date. Wikipedia later reproduces the 2015 date, citing the journalist's article.

Meanwhile, a government register records that the relevant legal entity was registered in 2017.

In addition, a corporate press release prepared by Company X has been posted on the website of a major news agency, in a dedicated section and without independent editorial selection.

This gives five documents.

Table 3. Hypothetical classification of Company X documents

Document	QISM level	Feature
Company X official website	V (Open Web)	Primary self-publication
Independent trade article	IV (Media)	Editorial paraphrase
Wikipedia article	III (Expert)	Derivative encyclopaedic document
Government registration entry	I (Government)	Official legal document
Corporate press release on the agency's website	V (Open Web)	Self-publication on a media platform

The example illustrates several properties of QISM. First, documents containing the same claim may belong to different levels. Second, the reputation of a platform does not automatically determine the status of a document published on it. Third, different primary documents may describe different circumstances.

The date on which a company began operating and the date on which its legal entity was registered are not necessarily mutually exclusive claims. The apparent discrepancy between 2015 and 2017 does not, without further analysis, establish an error.

In W3C PROV-DM terms, confirmed derivation relations can be expressed with *wasDerivedFrom*. If the news article did draw on the corporate document, and the encyclopaedia then cited the news article, the result is a chain of derivation.

The corporate press release may form an additional branch through which the same information spreads; its attribution to Company X is expressed with *wasAttributedTo*. No link between the press release and the corporate website has been established, but both documents come from the same producer. The four documents giving 2015 therefore all go back to a single source of information, Company X, and are not four independent confirmations.

The registration entry is an independent documentary source that relates to a different kind of fact.

The example shows that QISM can distinguish the document, the claim, the source of origin and the institutional level without replacing an investigation of the actual facts with a simple assessment of status.

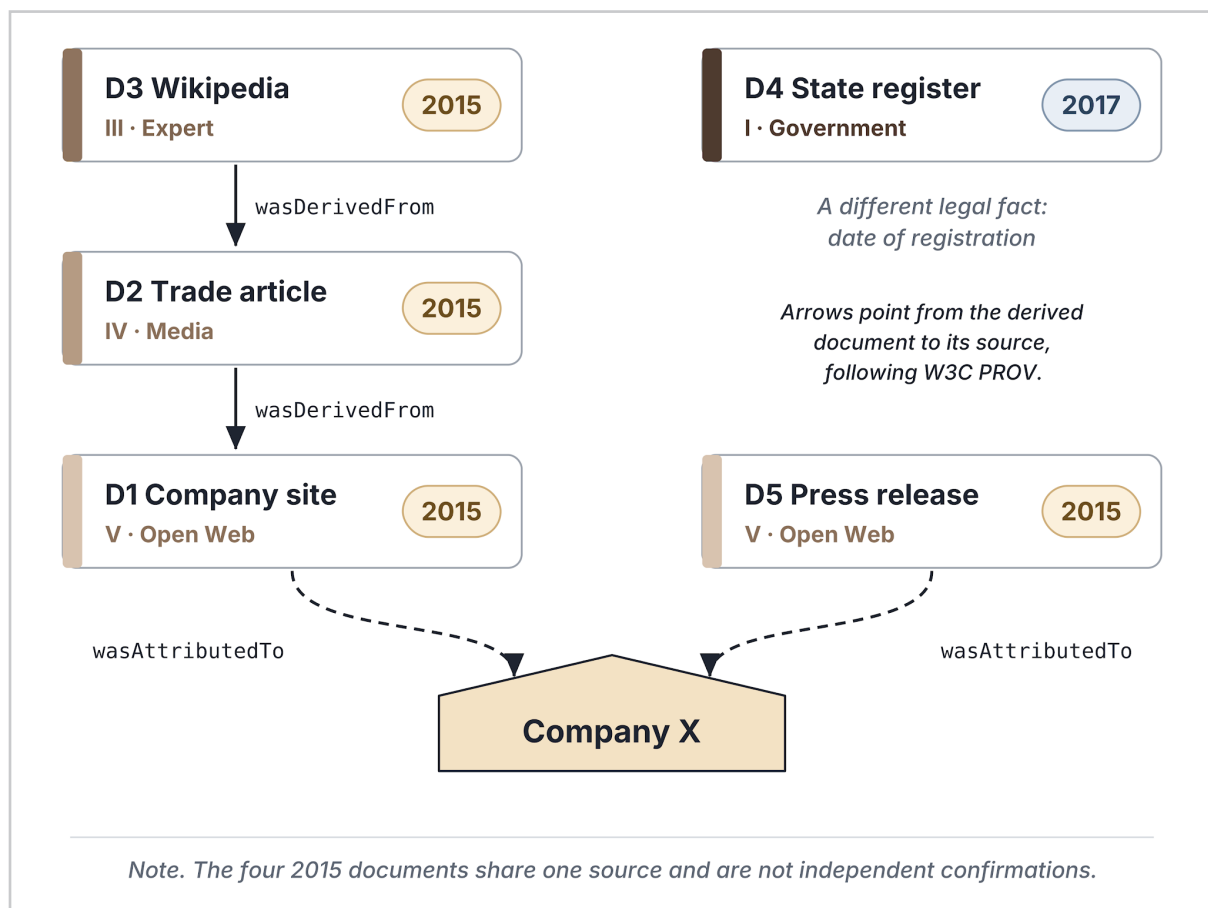


Figure 5. Hypothetical provenance diagram for the claims about Company X.

Discussion

The proposed model combines sociological and information-science approaches to studying public information about an entity.

Its distinctive feature is that it shifts the focus of research from the observable answer of a generative system to the structure of the information environment that exists before that answer is produced. AI Visibility can then be treated not only as a property of generative output but also as a possible outcome of the interaction between source information and the technical mechanisms that select and process it.

The properties of the public environment should not, however, be confused with the workings of the AI system itself. A generative answer may be shaped by training data, retrieval mechanisms, system instructions, the wording of queries, internal ranking algorithms, language and other factors.

Information stratigraphy therefore addresses one side of the research problem: describing the potential informational basis. The other side concerns the mechanisms by which AI systems obtain, process and reproduce information.

The nature of the proposed hierarchy also deserves attention. The five levels reflect differences in institutional procedure, but they do not assume that all verification mechanisms are fully comparable. The order of the levels should be treated as a working assumption that remains open to refinement.

Documents at the open-web level may turn out to be more visible in generative answers than official or academic publications. That would not necessarily refute the conceptual classification, because a document's institutional position and its actual presence in AI answers are different research variables.

This distinction opens the way to studying how social authority, information accessibility and algorithmic selection relate to one another.

QISM as the basis for a research programme

The scholarly significance of QISM goes beyond offering a new classification of public sources. The model sets out a sequential programme for studying the factors that may determine how organisations and people are represented in the answers of generative artificial intelligence.

1. The structure of the public information environment

The first line of research is the empirical description of the information environments of different entities.

Researchers can examine how documents are distributed across levels and sublayers, how companies differ from non-profit organisations, how language environments vary, and how far derivative publications depend on primary sources.

Such studies can refine the proposed stratigraphy and show how consistently independent researchers apply it.

2. Comparing information layers with AI Visibility

The second line of research involves measuring two structures in parallel: the public information environment of an entity and its representation in the answers of AI systems.

This makes it possible to test whether there are stable relationships between the distribution of documents across levels and the character of generative answers. One could examine, for example, whether a strong presence of independent expert documents is associated with how often certain information is reproduced.

The difference between how often a source is cited and how accurately its claims are conveyed deserves separate attention.

3. Information provenance

The third line of research concerns the provenance of specific claims.

Using the links between documents, a researcher can analyse which information moves between levels, where the wording changes and how errors spread.

A particularly interesting question is whether the likelihood that a claim appears in a generative answer changes as it is reproduced by several independent or dependent sources.

Counting publications is not enough for such a study. Independent confirmation of a fact has to

be distinguished from repeated copying of a single original report.

4. The institutional status of sources

The fourth line of research concerns the effect of a document's origin as such.

The experimental results of Dai et al. (2025) show that, under certain conditions, the identity of a media source can influence which source is cited. QISM allows the question to be extended to differences between institutional levels.

One can test whether the use of identical content changes depending on whether it appears in a government, academic, expert, media or open-web document. Such experiments need to control for content, format, accessibility, language and other variables.

5. The AI Access Pathway

The fifth line of research concerns the technical accessibility of information.

The existence of specialised APIs and data-supply agreements gives grounds for a hypothesis about differences between public platforms. One could examine whether material from platforms with contractual access channels appears in generative answers more often or more quickly than comparable material from platforms without them.

Such work must take account of the popularity of resources, search indexing, update frequency and the capabilities of the models themselves.

The interval between a document's publication and its appearance in the answers of search-enabled AI systems could become a subject of study in its own right.

6. Expert, statistical and experimental testing

Different research methods can be used to develop these lines of inquiry. Expert assessment can clarify the nature of institutional procedures and contested classification cases. Statistical studies can establish relationships between parameters of the information environment and observed AI Visibility. Controlled experiments can test the effect of particular changes to the content or provenance of documents while limiting alternative factors. Repeated measurement is needed to assess whether observed changes are stable.

It is especially important to distinguish correlational observations from causal conclusions. If a new publication coincides with a change in a generative answer, that is not enough to claim that the publication caused the change. Other sources, search indexing, model updates and the features of the queries must all be taken into account.

7. The sequence of research

QISM sets out the following research trajectory:

Describing the environment → measuring AI Visibility → comparison → formulating hypotheses → testing possible causes → repeated measurement and refinement of the model.

The first stage produces a structured description of the entity's public information environment. The second examines how the entity is represented in specific AI systems. The third identifies correspondences and discrepancies between the public information that exists and the generative answers observed.

Hypotheses are then formulated about the possible role of information levels, sources, claims and access channels, and subsequent studies test them by different methods.

Over time, such a programme can accumulate comparable results. These would help to refine the model and to identify the conditions under which the characteristics of the public information environment actually influence AI answers.

Limitations

QISM is conceptual and methodological in character. Its five-level structure and sublayers have not yet been tested empirically for reproducibility or completeness.

The model is confined to the public information environment. It does not directly examine closed training corpora, private data or the internal mechanisms of AI systems.

Using the document as the main unit of analysis needs further refinement for multifunctional platforms and mixed publication formats.

Information about contractual access channels does not, on its own, establish that specific documents have actually been used.

Future research should therefore test both the soundness of the classification itself and the presumed relationships between information levels and generative answers.

Conclusion

This article has proposed the QuorumLab Information Stratigraphy Model (QISM), a five-level heuristic model of an entity's public information environment. It rests on government, academic, expert, media and open-web levels, distinguished by the institutional procedures to which documents are subject.

Separating the document, the claim and the information producer allows the origin and spread of information to be studied as a research process in its own right. The additional AI Access Pathway dimension accounts for the potential technical and contractual channels through which AI systems reach public sources.

The immediate output of QISM is the ability to describe the information environment of an organisation or person systematically, map it, identify links between documents and establish where public information comes from.

Its broader scholarly value lies in providing a foundation for sequential research into the relationship between this environment and the answers of generative systems. QISM proposes a move from observing individual AI answers to studying the structure of available information, its institutional origin and the possible mechanisms of influence.

The model thus opens a distinct line of research in which the public information environment is treated not as a random collection of sources but as a structured object that can be classified, compared, assessed by experts and studied experimentally.

Declarations

Conflict of interest

Serhii Pontus is the founder of QuorumLab, and Leonid Kachuliak is its Chief Technology Officer. QuorumLab provides commercial services in the field of AI Visibility. QISM was developed as part of QuorumLab's research activity. This article presents a conceptual model and does not evaluate the effectiveness of the organisation's commercial services.

Funding

This research received no external funding. The work was carried out by researchers at QuorumLab.

Data availability. No original empirical dataset was generated for this study.

Use of artificial intelligence tools

AI-assisted writing tools were used in preparing the manuscript. The authors are responsible for the content, the scholarly argument and the final verification of sources.

References

- Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5–16). Association for Computing Machinery. <https://doi.org/10.1145/3637528.3671900>
- Dai, S., Cao, Z., Wang, W., Pang, L., Xu, J., Ng, S.-K., & Chua, T.-S. (2025). Media source matters more than content: Unveiling political bias in LLM-generated citations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 17256–17276). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.872>
- Entman, R. M. (2003). Cascading activation: Contesting the White House's frame after 9/11. *Political Communication*, 20(4), 415–432. <https://doi.org/10.1080/10584600390244176>
- Eysenbach, G. (2020). How to fight an infodemic: The four pillars of infodemic management. *Journal of Medical Internet Research*, 22(6), e21820. <https://doi.org/10.2196/21820>
- Google. (2024, February 22). *An expanded partnership with Reddit*. <https://blog.google/inside-google/company-announcements/expanded-reddit-partnership/>
- Hamann, A., Stahl, F., & Kannan, P. K. (2026). *Beyond consumer memory: How large language models update brand knowledge* [Working paper]. SSRN. <https://doi.org/10.2139/ssrn.6798479>
- Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, 44(4), 1467–1484. <https://doi.org/10.1016/j.ipm.2007.10.001>
- Hirvonen, N., Multas, A.-M., Nygård, T., & Huotari, M.-L. (2025). Cognitive authority: A scoping review of empirical research. An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, 76(1), 155–192. <https://doi.org/10.1002/asi.24942>

- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. *Proceedings of Machine Learning Research*, 202, 15696–15707. <https://proceedings.mlr.press/v202/kandpal23a.html>
- Kumar, P. (2026). *Generative engine optimization at scale: Measuring brand visibility across AI search engines* [Preprint]. arXiv. <https://arxiv.org/abs/2606.20065>
- Li, A., & Sinnamon, L. (2024). Generative AI search engines as arbiters of public knowledge: An audit of bias and authority. *Proceedings of the Association for Information Science and Technology*, 61(1), 205–217. <https://doi.org/10.1002/pra2.1021>
- Maltsev, O. (2023). *The screen: Product of the hyperreality machine*. Baudrillard Now. <https://baudrillard-scijournal.com/the-screen-product-of-the-hyperreality-machine/>
- Nguyen, P. A., Noorily, J., Flathers, M., Notsu, H., Ospina-Pinillos, L., Nguyen, T., Clark, S., Keane, A., Thompson, G., & Torous, J. (2026). *Sources of truth: A multi-platform, multilingual audit of citations in AI mental health information queries* [Preprint]. arXiv. <https://arxiv.org/abs/2609.00319>
- Reddit. (2024a, February 22). *Expanding our partnership with Google*. <https://redditinc.com/news/reddit-and-google-expand-partnership>
- Reddit. (2024b, May 16). *Reddit and OpenAI build partnership*. <https://redditinc.com/news/reddit-and-oai-partner>
- Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, 53(2), 145–161. <https://doi.org/10.1002/asi.10017>
- Seo, Y., Jeong, W., Kim, E., Jang, H., & Lee, D. (2026). *Verified misguidance: Measuring structural citation failures in search-augmented LLMs* [Preprint]. arXiv. <https://arxiv.org/abs/2605.28565>
- Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. Routledge.
- W3C. (2013). *PROV-DM: The PROV data model*. <https://www.w3.org/TR/prov-dm/>
- White, D. M. (1950). The «gate keeper»: A case study in the selection of news. *Journalism Quarterly*, 27(4), 383–390. <https://doi.org/10.1177/107769905002700403>
- Wikidata contributors. (n.d.-a). *Notability*. Wikidata. <https://www.wikidata.org/wiki/Wikidata:Notability>
- Wikidata contributors. (n.d.-b). *Verifiability* [Proposed guideline]. Wikidata. <https://www.wikidata.org/wiki/Wikidata:Verifiability>
- Wikimedia Enterprise. (2026a, January 15). *Announcing new Wikimedia Enterprise partners for Wikipedia's 25th birthday*. <https://enterprise.wikimedia.com/blog/wikipedia-25-enterprise-partners/>
- Wikimedia Enterprise. (2026b). *Wikidata API: Structured access to the world's largest open knowledge graph*. <https://enterprise.wikimedia.com/project-data/wikidata-api/>
- Wikipedia contributors. (n.d.-a). *Perennial sources*. Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources
- Wikipedia contributors. (n.d.-b). *Reliable sources*. Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources
- Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Greenwood Press.
- Žatuchin, D. (2026). *How large language models source brand reputation across languages and markets* [Preprint]. arXiv. <https://arxiv.org/abs/2606.25787>

Author Biographies

Serhii Pontus is the founder of QuorumLab, a European research startup specialising in AI Visibility. His professional background includes software development and information security. His research interests include public information environments, institutional classification of sources and information provenance.

Leonid Kachuliak is the Chief Technology Officer of QuorumLab. His professional background is in software engineering and technology development. He leads the technical development of QuorumLab's research infrastructure for studying AI Visibility.