Observatoire ARGA

INTERNATIONAL ANALYTICAL REPORT

Study No. 16

Independent Audit of Artificial Intelligence

Standards and Evidence-Based Methodology
Safeguards for Independence and Audit Quality

Organisation: Observatoire ARGA

English

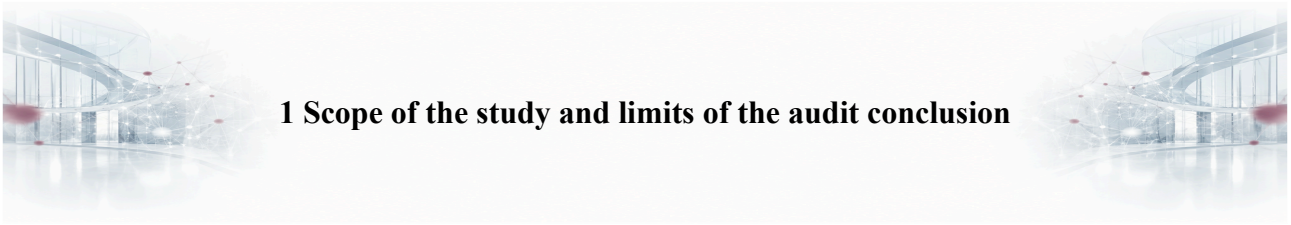
21 rue de l'Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

20 October 2026

Contents

1 Scope of the study and limits of the audit conclusion — 3
Scientific grounds and research design — 5
2 Regulatory and institutional framework — 7
3 European procedures and independence requirements — 9
4 Comparison of US and UK approaches — 12
5 Standards and selection of criteria — 14
6 Independence and conflicts of interest — 16
7 Team competence and quality control — 19
8 Terms of reference and conditions of access — 21
9 The audited system and the supply chain — 24
10 Provenance and preservation of evidence — 27
11 Data quality and provenance — 29
12 Sampling and test design — 31
13 Performance metrics and statistical uncertainty — 34
Summarized configuration comparison on file set — 36
14 Bias and implications for human rights — 37
15 Explainability and human oversight — 40
16 Auditing a generative legal assistant — 42
17 Robustness and information security — 45
18 Personal data and anonymisation — 47
19 Reproducibility of results — 50
20 Findings and the final audit opinion — 52
21 Monitoring and reassessment — 55
22 Application to ARGA's areas of work — 58
23 Regulatory gaps and international consultations — 60
24 ARGA's practical programme and conclusions — 63
Appendix 1 Limited audit task — 65
Appendix 2 Register of evidence — 67
Appendix 3 Conflict of interest assessment — 68
Appendix 4 Sampling plan and rare errors — 70
Appendix 5 Verification of the classification indicators — 72
Appendix 6 Differences between groups — 73
Appendix 7 Reproducible test report — 74
Appendix 8 Review of Legal Assistant's replies — 76
Appendix 9 Model limited conclusion — 78
Appendix 10 ARGA package of proposals — 79
Index of Practical Recommendations — 81
References and Supporting Sources — 82



1 Scope of the study and limits of the audit conclusion

What issue should the audit resolve

An independent audit of artificial intelligence makes practical sense when it allows you to verify a certain statement about a system and name the basis for the conclusion. The subject of the check may be compliance, process quality, test result or effectiveness of the protection measure. The formula "artificial intelligence has been audited" without specifying the object, version, criteria and limitations, does not indicate exactly what it is.

The research question is formulated in an operational way: under what conditions can an independent researcher reproduce a conclusion about a specific legally significant property of AI and determine the permissible area of its use? Four related prerequisites are verified: the property is correctly determined; observations allow it to be measured; the conclusion is resilient to reasonable evaluation changes; the decision procedure is protected from conflict of interest. The source comparison provides a basis for the proposed methodology; its empirical effectiveness is subject to separate verification by the protocol.

For ARGA, the subject is particularly important in legal files, name matches, financial analysis and text preparation. An error can change a person's assessment or the circumstances of the case. Therefore, the proposed approach links technical verification with the consequences for the affected person: the availability of correction, the clarity of reasons and the possibility of restoring the decision making process.

In the report, the word "audit" is used as a generic reference to systematic verification against established criteria. This working definition does not confer on the executor the status of a notified body, accredited laboratory or certification body. The legal value of a particular conclusion is determined by the applicable procedure. The separation of such procedures is the starting point for the entire proposed methodology.

Task Element	What needs to be determined
Object	System, version and specific application
Approval	Controlled property or performance of duty
Criterion	Rule, contract clause or approved methodology
User of result	Who and for what decision receives the opinion

Which is considered proof.

The evidence is assessed with respect to the assertion and possible alternative explanation. If the answer contains a correct quote, it confirms the existence of the relevant fragment, but it does not yet confirm the completeness of the answer or the applicability of the norm. If the log shows a

reference to a document, it remains to see if it affected the result. For each substantive statement, therefore, the primary object, the operation performed, the observation received and the limit of interpretation are recorded. This entry makes it possible to distinguish direct confirmation from an indirect sign.

It is suggested to distinguish evidence about the structure of the system, its behavior and the consequences of its application. A version passport answers a question about the object; a replicated test answers a question about the behavior being observed; the material of the actual process answers a question about the consequences. Separate justification is required between levels. Successful testing of an artificial file does not prove a reduction in the number of procedural errors in practice, and the established harm does not allow for a cause to be attributed to a particular component without investigation.

The verification shall be based on the approval to the procedure. If it is declared that the documents have not been transferred to external services, it is not enough to read the contract. You'll need to explore the routes of transmission, settings, permissions, and observed behavior in the permissible environment. If the accuracy of identification of persons is stated, a legal examination of the conditions of application will not replace the test on appropriate data.

The negative result is of particular importance. The absence of a found incident, the absence of access to information and the confirmed absence of a specific action are different conditions. They are reflected separately in the detention. Otherwise, the lack of evidence turns into an unfounded confirmation of well-being, and the customer gets a conclusion that the work actually does not support.

Material	Supported conclusion	Significant limitation
Supply contract	Agreed commitment	Doesn't prove execution
log of events	Recorded transactions	Need to check completeness
Check Start	Observed behaviour	Limited to launch conditions
Staff interview	Understanding and Process Description	Requires comparison with practice

Method of study and status of proposals

The legal part is based on official EU, U.S. and British publications. The ISO standards are reviewed by means of open object and editorial descriptions. Full paid texts are not used for the reconstruction of mandatory paragraphs. Therefore, the proposed forms of verification cannot be given for a complete list of requirements of any standard or a ready-made certification audit programme.

The normative map reflects the status of the sources as at 20 October 2026. The applicability of the standard is determined separately from its publication: the territory required, the role of the participant, the category of the system, the date and the exceptions provided. Advisory projects are identified as projects. In particular, the proposed EDPB guidance on anonymisation and data collection for generative AI is not presented as a final rule. [27], [28], [30], [31]

The methodological decisions of the report are proposals for the organization of audits. They include a record of evidence, a test report, an independence assessment and a form of conclusion. In each practical example, conditional circumstances and educational numbers are used. The report does not contain the test results of the actual supplier and does not confirm that ARGA has official valuer status.

When applying the methodology, the permissible amount of work is selected first. One set of competences is sufficient for documentary verification; technical evaluation of the model requires another. If a substantial issue is beyond the team's capabilities, it's passed on to a professional or directly restricted. This restriction protects the quality of the research and allows the customer to understand what additional work needs to be ordered.

The practical result should answer three independent questions: what is established, how reliably it is established, and what action follows from the result. Neither a positive evaluation of the organizational process nor a successful test of a particular function exempts the user from determining the legality of a particular application of the system.

Scientific grounds and research design

The scientific directions are used as the basis for individual research operations. Measurement theory links the concept of legal error to the signs of response; statistics defines the unit of comparison and uncertainty; software engineering provides configuration identification and re-analysis; reporting research links the finding to responsible action. The procedure below adapts these grounds to the legal assistant and is subject to pilot inspection. [M1–M3]

Scientific direction	Operation in the study	Test result
Measurement theory	Definition of significant error and marking rules	Classifier with positive and border examples
Applied statistics	File-level couples design and uncertainty analysis	Reference quantities and reproducible calculation of the difference
Programme engineering	Recording of the model, search, instructions and rights of the tools	Passport configuration and playbook
Algorithmic Accountability	Relationship of finding to correction and responsible decision	Motivated solution for each significant defect
Legal methodology	Checking the applicability of the standard and the consequences of an error	Separate factual, technical and legal findings

Research question and analysis unit

The pilot's main question is: Does the configuration under investigation change the file's portion with at least one significant error in keeping the required data complete? The object is the entire configuration of the assistant, including extracting documents and limiting actions. The primary analysis unit is a file. Detailed labels of individual claims explain the mechanism of the defect, but do not automatically increase the number of independent observations.

The definition of error must be linked to the intended application. For internal chronology, a date or source change is essential; for the preparation of a legal position - loss of exclusion, wrong

jurisdiction or change in the status of a fact. It is suggested that a map should be drawn in advance of the "species of error - possible consequence - the way of detection". Scientific validity here means the validity of measuring a particular property, not the attribution of a general quality score to the answer. [M1]

The hypothesis and the plan for comparative testing

For the main comparison, the original configuration A, modified B and the same files are specified. The null hypothesis refers to the absence of a directed difference in the pair outcomes; the research hypothesis refers to a reduction in the proportion of significant errors. Practical reduction and unacceptable new defects shall be determined before starting. Even statistically distinct results may not justify implementation if the change is too small or makes the critical category worse.

The order of presentation of tasks and answers is mixed; the experts, if possible, hide the appropriateness of the configuration response. All scheduled runs, failures and technical failures are maintained. For multiple launches on one file, the regulation of aggregation or model of observation is selected in advance. Separately report the quality on the fixed set and the basis for transferring to potential new tasks. [M3]

Verification of measurement tool

Before the main assessment, two experts independently mark the training material. The discussion is used to clarify the instruction, after which its suitability is checked on a new subset. The usefulness of markings is assessed by specific categories of errors and reasons for discrepancies. Achieving consensus after discussion does not replace the measurement of the original reproducibility; the contentious case remains as the observed limitation of the reference.

Competing explanations for improvement are checked: shorter answers, increased rejections, elimination of complex files, or a change in the search engine. So next to the main outcome is the completeness of the compulsory facts, the proportion of correct failures, and the distribution of types of tasks. If more than one component is changed at the same time, the result relates to the sum of the components. A separate comparison is required to determine the contribution of one component.

Conditions for acceptance and revision of the finding

Restricted access is possible after data checks, calculations and significant new defects. The conclusion indicates the area under study, the remaining uncertainties and changes that require a new assessment. These include changing the model, the search engine, the language, the power of the tool and the purpose of the answer. The protocol allows the decision to be taken "not enough" and lists the specific information or tests required for the following decision.

The result of this report is a draft methodology and tested design. The hypothetical calculations below demonstrate the application of the rules, but do not replace empirical research. Scientific testing of the methodology itself should assess the reproducibility of expert conclusions and its ability to detect legally significant defects with limited resources. This division allows you to formulate a verifiable research task for the next pilot.

Practical case analysis 16.01

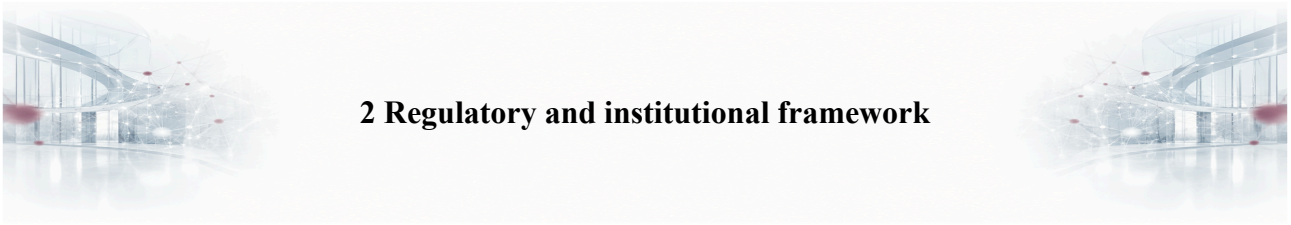
Recommendation: Record the narrow, verifiable approval

Hypothetical example: The commissioner asks for confirmation of the reliability of the assistant, who makes up the timeline of the case.

Analysis: The correct date in one answer does not confirm the completeness of the timeline. The date, source and missing events are checked separately.

Practical steps: The audit manager approves the list of compulsory events and compares the answers to the original documents.

Outcome criterion: The conclusion relates to the preservation of events and sources in the research kit.



2 Regulatory and institutional framework

Different grounds for different checks

The obligation to carry out an inspection may arise from a law, a supervisory requirement in a given field, a contract or a voluntary programme. These grounds differ by addressee, content and effect. To start the work, you need a validity record: who is obliged to act, which object is covered, when the obligation comes and what document should be obtained.

In the EU, the Artificial Intelligence Regulation provides for several conformity assessment procedures. The participation of an external notified body depends on the system category and the intended route. The high-risk attribution alone does not suggest that any supplier is obliged to order a universal independent audit from an arbitrary commercial organization. [3]

Individual regimes operate outside this regulation. The European Digital Services regime provides for independent annual audits of particularly large platforms and search engines. In New York, a special audit of certain automated personnel decision-making tools has been set up. None of these designs apply to any AI just because it uses an algorithm. [6–9]

Basis	What is installed prior to verification
Mandatory procedure	Addresser, competent authority and authorized performer
Treaty verification	Agreed criteria, access and impact of the result
Voluntary evaluation	Purpose, method and limits of the confirmation declared
Internal audit	Accountability of the performer and how the deficiencies are addressed

For ARGA, this card allows you to offer a useful service within your competence: document research, legal characterisation and verification of the consistency of the claims. If the customer needs a document of a legal form, it is first established who has the right to issue it. The substitution of such a document with an analytical opinion poses a risk to both sides.

Date of rule and date of duty

The verification calendar shall take into account the relevant transitional provisions. Article 113 of the EU AI Regulation retains the general date of application of 2 August 2026 and the special dates. For sections 1-3 of Chapter III, except for Article 6(5), 2 December 2027 for systems under Article 6(2) and Annex III and 2 August 2028 for systems under Article 6(1) and Annex I are indicated. [1]

It cannot be inferred from this that any data processing or any AI solution is permissible before these dates. Data protection rules, labour, consumer and sectoral legislation may also apply. Checking readiness for future requirements and assessing compliance with already applicable norms have different subject matter and should give different formulas for the outcome.

It is useful to record the date not only at the legal act but also at the technical standard. ISO 19011 is published in the new version of 2026. If the agreement continues to refer to the 2018 edition, the auditor must determine whether the parties deliberately chose it and whether a change of assignment is required. The replacement of the version of the standard shall not occur undetected to the opinion addressee. [16]

Recording in calendar	Audit significance
Date of validity	Determines legal assessment of behaviour
Start date of the compulsory procedure	Determines the required confirmation route
Revision of the criterion	Allows the verification content to be restored
Date of test	Limits the technical relevance of the result
Change of configuration	May require a review of an earlier finding

The working register should keep the past records. Otherwise, a late adjustment to the deadline would destroy the explanation as to why the previous decision was considered justified. For each change, it is enough to indicate the official source, the date of review and the assignment items concerned.

Who makes the decisions after the audit

The auditor shall establish the facts and evaluate them according to the criteria chosen. The head of the audited organization shall determine the measures within the limits of his or her authority. The regulator, court or other authorized body shall retain its own jurisdiction. The transmission of the report to an external addressee does not make the auditor's opinion a binding decision for the addressee.

In the proposed order, four roles are allocated: process owner, verifier, independent reviewer and decision maker. A single employee may combine some functions in a small organization, but must not prove the independence of the assessment of his own dispute resolution. If separation is not possible, the restriction is disclosed and the substantive conclusions are passed on to an outside reviewer.

It is particularly dangerous to dilute responsibility after the involvement of the supplier. The customer can assume that the test was completely carried out by the technical team and that the legal admissibility was confirmed by the lawyer. Therefore, each substantive assertion is linked to a specific perpetrator and a specific procedure. The general signature of the project manager does not replace such a connection.

For socially significant systems, a channel for obtaining objections from affected persons should be provided. A user's request may not always prove a system defect, but it may show a situation that did not fit into the test sample. The audit program should allow such a evidence to be examined without automatically revealing a personal file to the entire team.

The outcome document indicates to whom the conclusion is addressed and which decisions it supports. For example, confirming the completeness of logging may be enough for the next stage of testing, but not enough to run a system that makes meaningful decisions about humans. The sequence of decisions should be clear before the inspection begins, not after the report is issued.

Practical case analysis 16.02

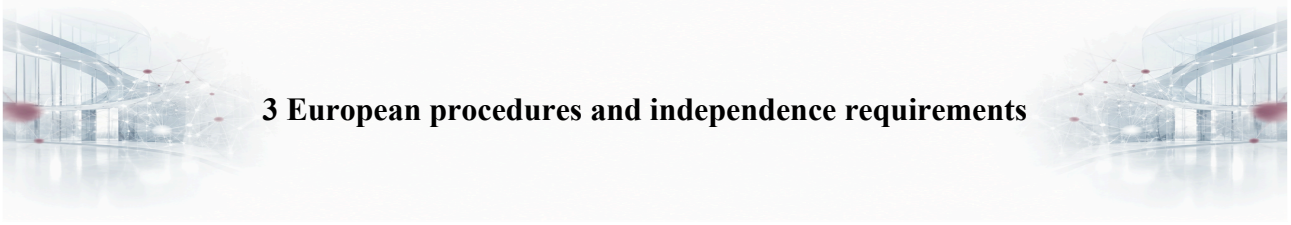
Recommendation: Make a map of applicable requirements

Hypothetical example: The supplier offers one certificate for bank screening and personnel selection.

Analysis: The same model does not mean the same responsibilities. The designation, entity and territory define the individual verification issues.

Practical steps: The lawyer describes the two applications and for each, indicates the basis, the address of the duty, and the required evidence.

Outcome criterion: Each requirement relates to a specific application and source.



3 European procedures and independence requirements

Notified body and normal external auditor

Article 31 of the EU AI Regulation sets out special requirements for notified bodies. They shall be independent of the verified supplier and other economically interested operators. Organized measures of impartiality are provided for; participation in development and certain advisory services may conflict with the evaluation function. These are specific institutional status requirements. [2]

An external consultant performing a contractual verification does not become a notified body due to the registration of a legal entity, the presence of experienced staff or the use of an international standard. The procedure provided for by law is necessary for the relevant status. The contract and the public description of the service must contain names that accurately reflect the authority of the executor.

A specific standard can provide a methodological guide for voluntary verification, but its legal effect cannot be transferred to all market participants. The restrictions on the combination of consulting and audit proposed in the report are applied as internal quality guarantees unless

otherwise required. This distinction allows for greater independence without creating a universal legislative obligation.

Question under review	Sufficient confirmation
Official status declared	Registration and scope of authority in the applicable system
Technical competence declared	Composition and proven experience
Independence declared	Verification of relationships and protection measures in place
Certification declared	Scheme, object, body and area of certificate

For ARGA, the correct starting position is to describe the actual research or expert work being done. The possible development of a more formal function requires a separate analysis of competence, resources, responsibilities and permissible status. The present report does not declare such a function already established.

Conformity assessment and technical case files

Article 43 distinguishes between internal control procedures and procedures involving a notified body. For the groups of systems it provides, sectoral legislation, the application of harmonised standards and general specifications are important. Appendix VII links the assessment of the quality management system and technical documentation with the established observation. [3–4]

For voluntary auditing, the principle of separating an object into a process and a specific technical implementation is useful. The organization may have a detailed description of the release order, but there is no proof of its execution for the verified version. The reverse is also possible: a single product is well-tested, and change management remains unstable. These results should not be reduced to one non-differentiated score.

It is proposed to request two related sets of materials. The first shows the rules of governance: allocating responsibility, approving changes and handling inconsistencies. The second shows the implementation: versions, test logs, data composition and confirmation of specific defects. There should be links between sets to restore the path from a rule to an action.

Organizational document	Implementation to be checked
Issue order	Decision to issue a certain version
Retesting requirement	Protocol after significant change
Procedure for correcting errors	Recording the defect and confirmation of its elimination
Rule on admission of staff	Real access rights and history

If part of the material is unavailable for an objective reason, the auditor considers alternative procedures. Surveillance in a protected environment sometimes gives more information than an impersonal hold. However, the alternative procedure must actually answer the question; it is not used as a formal way of circumventing the lack of evidence.

Evidence of Digital Platforms

The European Commission explains that particularly large online platforms and search engines under the digital services regime are subject to independent audits at least once a year. The delegated regulation 2024/436 specifies the procedure for such checks and reporting. It's a separate design of service-specific control that can't be represented as a general rule of auditing all AI models. [6–7]

Its significance for this study is the relationship between verification, documented conclusion and subsequent actions of the verified party. A similar traceability is offered for the ARGA voluntary program: a significant disadvantage is provided by the owner, substantiated correction and separate verification of the result. The very promise to eliminate the problem does not close the door.

Publicity requires self-employment. A report for guidance may contain technical information, personal data and a description of vulnerabilities that cannot be transferred to open access without risk assessment. At the same time, a public resume should not exclude substantial restrictions so that the reader gets the opposite impression of the results of the review.

A full report, a protected application and a public summary are prepared in the proposed order. For each seized part, the reason for the restriction of access is recorded. The reader is informed of the subject, period, types of procedures, substantive results and limits of the conclusion. The details of the attack can be limited by keeping a message about an unresolved critical flaw.

Such a device does not create the right to publish confidential information to others. The conditions for the use of the report are agreed in advance and are in accordance with the law. If the customer requests an advertising declaration that is inconsistent with the established results, the auditor should be able to refuse to use his name and conclusions appropriately.

Practical case analysis 16.03

Recommendation: Check the authority of the external evaluator

Hypothetical example: The manufacturer shows the certificate of its quality system instead of confirming the authority for the required procedure.

Analysis: The evidence of organisation does not prove competence in a particular subject or route of assessment.

Practical steps: The purchaser asks for the area of authority and compares it to the task before the contract is signed.

Outcome criterion: The contract specifies the permissible type of detention without extending the status of the perpetrator.



4 Comparison of US and UK approaches

New York as an example of narrow mandatory verification

New York's rules on individual automated HR tools link their use to bias auditing, reporting, and notification. The official page indicates an audit within a year prior to use. The sphere is defined by the established concept of the tool and the conditions of use, not any automation of the personnel function. [8]

The final rules reveal the auditor's independence: participation in the use, development or dissemination of the instrument, labor relations and certain financial interests prevent recognition of independence. They also provide for calculations by established categories and requirements for the original data. These provisions cannot be replaced by a general declaration of professional objectivity. [9]

The methodological conclusion is to check the team's connections to the tool itself. Organizational distance from the employer still does not exclude the auditor's participation in the development of the supplier's product. Verification requires a history of relationships, not just a current position. The form proposed in the annex covers direct and indirect communications, subcontracting and payment dependence on the outcome.

Narrow result	Why doesn't he give general permission?
Calculation of differences between groups	Does not establish the full legality of employment
No margin identified	Doesn't rule out the error of selection criteria
Publication of summary results	Failure to confirm the completeness of all checks
Special audit	Does not waive other applicable obligations

This experience is being used for comparison outside the jurisdiction concerned. If an international organization voluntarily uses similar indicators, the task must explain the reasons for choosing the categories and the lawfulness of processing the required data without claiming that New York rules have become binding on it.

American methodological and oversight tools

The 2021 GAO U.S. Accounts Chamber's accountability framework is organized around management, data, performance and monitoring. It contains questions that are useful for verifying the use of AI by government organizations and other entities. It's a methodological source for building a program, not a universal scheme for letting private auditors into the market. [12]

The revised U.S. Risk Model Guide for 2026 has an independent banking context. Its usefulness for comparative analysis lies in its attention to model designation, verification, and limitation management. Generative and agent models are explicitly excluded from the scope of this manual. It cannot be presented as a special mandatory regime for their validation; using separate approaches outside the established sphere is an independent methodical solution. [13]

For ARGA, it is suggested that these materials be used to verify the completeness of questions: who owns the process, how useful the data is, what measures the test, and how the deterioration is detected. Each issue is transformed into a specific procedure and the expected evidence. You can't just state that the frame fits because four of its topics are listed in the report.

Different meanings of the same words are particularly important in the international draft. Model validation, organization audit and legal evaluation have different subjects. If a supplier sends a validation report, the auditor should read the actual volume of actions performed: whether there were independent data, whether application restrictions were checked and whether the real consequences of errors were considered.

The comparison shows the risk of false interchangeability: a detailed examination of one framework may leave an important question in another jurisdiction unanswered. Instead of a general declaration of international conformity, it is better to draw up a map of transferable evidence and specify additional procedures for a particular addressee.

The British Reliability Market

British materials on AI assurance look at different ways to gain valid confidence in the properties of systems. The Independent External Assessment Roadmap, published on 3 September 2025, is devoted to the development of quality services, competencies and markets. It's a software document; publishing it doesn't mean introducing a single mandatory certification of any AI. [10–11]

The choice of service is essential for the customer. Documentation checks, impact assessments, technical testing and certification answer different questions. If the task is to verify the correctness of the processing of legal texts, an assessment of the maturity of management alone will not provide an adequate answer. On the contrary, a large set of test queries will not show who is responsible for correcting errors after implementation.

Approach	A useful result	Transfer limit
EU and special procedures	Set form for the evaluation of a particular object	Enforceability and proper performance required
New York and human resources tools	Narrow verification of bias by local rules	Does not cover the full validity of the decision
US methodological framework	Organization of management and testing	Not a general operating permit
British policy participation	Development of quality external services	Does not replace sectoral responsibilities

For ARGA, it is suggested to choose a work model based on the needs of the client rather than the most compelling title. The subject matter, the qualifications of the performers, the availability and type of the final document shall be determined jointly. When multiple types of verification are required, their results can be combined, keeping separate grounds and limitations for each conclusion.

Practical case analysis 16.04

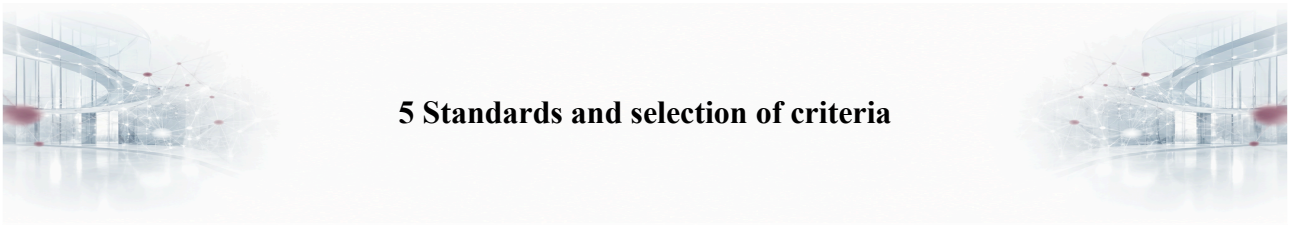
Recommendation: Segregate jurisdiction and appointment of verification

Hypothetical example: A human resources instrument report from one jurisdiction is offered to a bank in another country.

Analysis: The object, criteria and effects of decisions differ. Postponement of the result requires self-justification.

Practical steps: The lawyer draws up a table of matching and missing criteria; the technical expert assigns additional tests.

Outcome criterion: For each transferable conclusion, a basis or restriction is indicated.



5 Standards and selection of criteria

Current revision and audit area

The official ISO 19011 record indicates the 2026 edition and the subject of management system audit guidance, including program management and auditors' competence. The standard itself is not an independent certification scheme. For the present report, it serves as a guideline for setting an audit program; specific procedures are further proposed as author methodology. [16]

ISO/IEC 42001:2023 is a management system of the IE in the organization. ISO/IEC 42006:2025 describes additional requirements for bodies conducting audits and certification of such systems and complements ISO/IEC 17021-1. These documents have different addresses according to their public descriptions. An organizational system certificate cannot be automatically presented as proof of the integrity of each product. [17–18]

When the contract is concluded, the standard designation, the drafting and the source of its obligation for the assignment are recorded. If the parties agree to comply with specific paragraphs, the team needs legal access to the full text. An open record allows you to verify the name and subject, but it does not replace the professional application of all document requirements.

Document	The object of the public description
ISO 19011:2026	Management Systems Audit Manual
ISO/IEC 42001:2023	Management system of the organization

Document	The object of the public description
ISO/IEC 42006:2025	Audit and certification bodies for such systems
ISO/IEC 23894:2023	IA Risk Management Manual

The mission should not contain a promise of compliance with "all international standards". The test kit is always limited. If the criteria change during the work, the parties document the consequences: what procedures should be repeated, what data should be added and what conclusions should not be made in the previous formulation.

Test and confirmation of approvals

ISO/IEC 17025:2017 relates to the competence of test and calibration laboratories, and ISO/IEC 17029:2019 - validation and verification bodies. Their open descriptions show that different types of confirmation require different organizational and methodological grounds. The existence of a single status does not prove the ability to perform any I.E. verification. [19–20]

To choose an performer, it's helpful to ask for a proven area of competence and compare it with a real task. A laboratory can measure a certain technical indicator well, but it cannot analyze the lawfulness of processing personal data. The legal team can set responsibilities but not be prepared for statistical design of the test. The written distribution of these tasks prevents hidden gaps.

In a voluntary programme, it is necessary to determine separately whether an affirmation of an action already carried out is confirmed or the justification of the intended application is assessed. A past success in one environment doesn't guarantee future use in another. The auditor indicates which data relate to the observed experience and which to the prediction or conditional assumption.

It is proposed to draw up each criterion as a short record: the verifiable endorsement, the basis, the procedure, the expected material and the permissible limit for the conclusion. This makes it apparent that there is a difference between fact and evaluation. For example, "there is a magazine" is checked differently from "the magazine allows every significant activity to be investigated". The second question requires an examination of the content, completeness and accessibility of the record.

The benefit of the standard arises from its precise inclusion in such a programme. The list of numbers on the title sheet does not replace the choice of methods and does not confirm that a particular test was performed by a competent and independent party.

Risk management and impact assessment

ISO/IEC 23894:2023 is dedicated to the management of AI risks, and ISO/IEC 42005:2025 is dedicated to the assessment of the impact of AI systems. In open descriptions, these are complementary directions: the organization explores uncertainty and possible consequences for people and society. The report does not reproduce the unaffordable full requirements of these standards. [21–22]

The NIST AI RMF 1.0 framework is voluntary and organizes work through management, context-determination, risk measurement and handling functions. It does not set a universal acceptable level

of risk. For an audit, this means finding out who has chosen the eligibility criterion and what justifies it in a particular area. [14]

It is proposed to separate measurement from relevance. Not every significant harm is easy to express in a single number. The disclosure of a confidential file, the inability to dispute an erroneous conclusion or the use of unacceptable grounds may require a high-quality legal assessment. The low frequency of such events does not automatically make them acceptable.

Evaluation issue	Output required
Who could get hurt?	List of affected groups and situations
How harm occurs	Description of causal sequence
What is measured	Indicator with area of applicability
Who accepts residual risk	The decision of the authorized person on the ground
When the conclusion is reviewed	Event or change starting the check

The auditor checks the quality of this reasoning and the evidence. He should not inadvertently replace an organization's decision with his own preference or take any of its decisions as sufficient. If the admission criterion is contrary to the applicable norm, the consent of the management does not correct the legal defect.

Practical case analysis 16.05

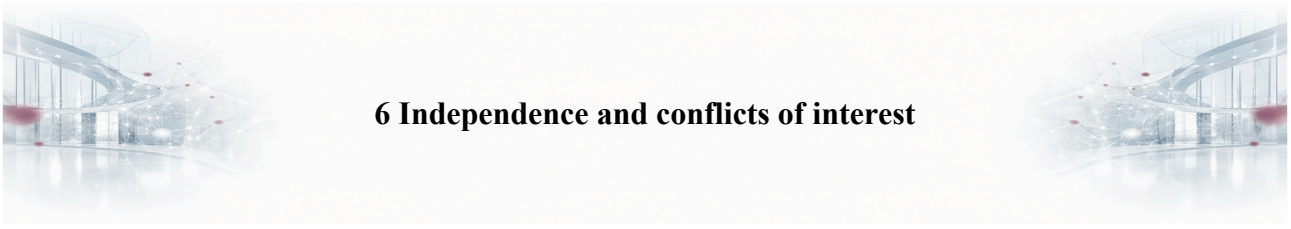
Recommendation: Select the criterion before testing

Hypothetical example: The supplier, after the test, selects an indicator by which his product looks better.

Analysis: Such a decision conceals bad outcomes and makes the evaluation dependent on the outcome.

Practical steps: The auditor records a significant error, the completeness of the response, the refusal and the rule of admission in advance; records changes.

Outcome criterion: The outcome contains all pre-selected indicators, including unfavourable ones.



6 Independence and conflicts of interest

From application to relationship verification

Independence involves the ability to choose the necessary procedures, evaluate materials and report inconvenient results without unacceptable influence. The external legal address alone does not ensure this. The verification may be limited by financial dependence, prior participation in the project, personal connections, or the control of the customer over the publication of the conclusion.

The proposed program examines the relationship between the organization, the project manager, the technical executives and the reviewer and the system being tested before taking a task. Special attention is paid to product design, setting, selling and escorting. If the employee has previously made a dispute resolution, the subsequent assessment of that decision creates a risk of verifying his or her own work.

Threat	Fact checked	Possible action
Self-interest	Share in business or dependent payment	Refusal or elimination of interest
Checking your own work	Participation in the development of the audit function	Other performer and separate reviewer
Protection of client position	Public promotion of the disputed product	Division of roles and limitation of task
Pressure	Threat of payment or access	Escalation and termination of employment

These categories are used as a methodological map, not as an affirmation of a single international list of prohibitions. When there is a mandatory procedure, its special rules apply first. In a voluntary review, the decision should take into account the significance of the threat and the real effectiveness of the safeguards.

A staff member's declaration alone is not enough for a complex project. It's compared to the contracts, team composition, and history of work in the available volume. A hidden conflict discovered later requires a review of the procedures involved, even if the original conclusion has already been issued.

Payment and control of content

Payment by the customer is the usual economic model of external verification, but the contract should not turn a positive conclusion into a reward condition. It is proposed that the payment be linked to the phases completed and the agreed scope of procedures. A negative finding is not considered a service deficiency if it is substantiated and received within the scope of the assignment.

The customer's right to correct an actual inaccuracy must be distinguished from the right to remove an unfavorable conclusion. Before the report is completed, the party being examined has the opportunity to submit documents and objections. The auditor shall record the consideration of each substantive objection. The opinion is amended for evidence, not as a result of negotiations on reputational convenience.

If one group simultaneously sells the implementation and verification of the implemented system, the internal team division doesn't always eliminate the threat. We need to evaluate the overall financial interest, leadership, awards and material sharing. When there is a significant risk, the most convincing solution may be to involve another organization to verify certain claims.

Condition of the treaty	Proposed protection
Method of payment	Remuneration is not dependent on a positive conclusion
Access to materials	Right to request the necessary information and record refusal
Harmonization of project	Correcting the facts without the right to hide the evidence
Use of name	Prohibition of misleading advertising
End of mission	Documentation of causes and significant risks already identified

The disclosure of the terms of payment is determined by law and contract. Even if the amounts are confidential, the final document can tell who ordered the check and what measures protected independence without giving the reader a false impression of the origin of the funding.

Verification of independence during the project

Independence is not assessed only when signing a treaty. In the process, team members, payment terms, supplier ownership or the nature of counselling may change. It is therefore recommended that the check be repeated in case of substantial change and before the conclusion is signed. The new conflict is considered before its influence is embedded in the final outcome.

A conflict resolution shall contain the circumstances, the tasks involved, the protection measure and the person who has assessed its suitability. The "conflict resolution" formula does not make it possible to understand what happened. For example, replacing a reviewer eliminates his or her personal connection, but it doesn't necessarily fix the test he or she has already done.

For a small project, a proportional procedure is possible: a short declaration, verification of the obvious relationship and independent consideration of the substantive issue. A complex project with sensitive implications would require a more detailed assessment. The scope of the document should not change the result: it is important to determine whether the perpetrator is able to draw an unfavorable conclusion and to preserve its justification.

If independence is not possible, let's do another kind of work, such as consultation analysis with an open description of the role. However, renaming should not hide the same problem from the user of the result. You can't give an advice report for an independent confirmation and then refer to a small caveat in the appendix.

The practical criterion is simple: the external reader must understand who tested the system, what connections could have influenced the assessment and why the measures taken are considered sufficient. If such an explanation is not possible without default, the independent status of the conclusion is unfounded.

Practical case analysis 16.06

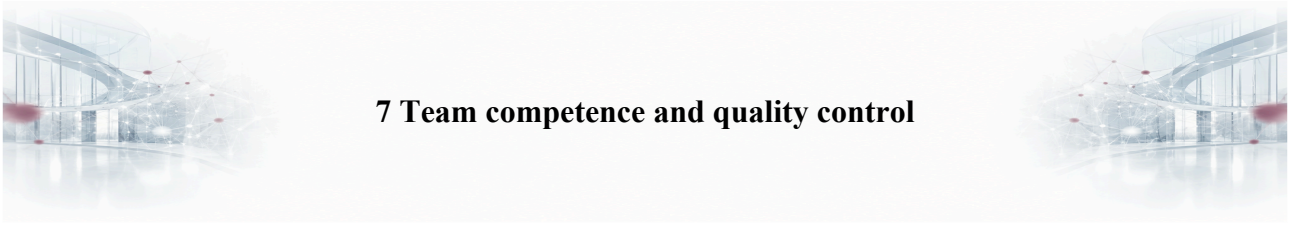
Recommendation: Eliminate conflict of interest before appointment

Hypothetical example: The lead auditor had previously designed an audited filtering mechanism.

Analysis: It evaluates its own decision; the declaration does not in itself eliminate dependence.

Practical steps: The coordinator will transfer the inspection of the mechanism to an independent specialist and retain the decision to withdraw.

Outcome criterion: The author of the mechanism does not support the conclusion that it is effective.



7 Team competence and quality control

Competences to be tested

The team is formed after the target is identified, not the other way around. The legal assistant may need legal expertise, knowledge of document processing, statistics and information security. The composition will be different for the image recognition system. The overall experience with AI does not confirm competence in all these areas.

It is proposed to separate the mandatory skills from the supporting skills of the project. The manager is responsible for the consistency of the program, but is not obliged to perform each specialized procedure independently. In doing so, he must understand the limits of the conclusions and be able to see when a technical answer does not solve a legal question. A contractor with special knowledge receives a clearly defined assignment.

Role	Tested Skills	Project confirmation
Head	Link objective, methods and conclusion	Approved programme and decisions on restrictions
Legal officer	Establish applicable requirements	Motivated Responsibilities Map
Technical specialist	Play the observed behavior	Protocol and Saved Results
Data specialist	Based on sample and indicators	Calculation and analysis of uncertainty
Reviewer	Test the logic of the conclusion	Comments and their authorization

The qualification statement is evaluated in relation to a particular task. The certificate of education may show the completion of the course, but not the completion of the comparable examinations. Experience should be confirmed by acceptable examples of work, description of the role and ability to explain the method used. The confidentiality of previous projects does not justify an unfounded promise of competence.

If the right specialist is not available, the program changes before testing begins. A significant gap cannot be hidden by the general formula of an interdisciplinary approach. A limited, but correctly executed check is more useful than a broad conclusion for which the team lacks the necessary knowledge.

Independent review

The content review checks the output-to-data chain. The reviewer receives the protocol, the original numbers, the error classifier and examples of the disputed markup; then independently restores the main calculation and chooses at least one plausible alternative interpretation. The result of the review is a list of verified statements, substantial objections and conditions under which the conclusion remains substantiated. Correction of style does not by itself fulfil this function.

It is proposed to highlight the mandatory subjects of the review: the applicability of the criteria, the completeness of significant procedures, calculations, contradictory data and the formulation of the final opinion. Independent verification of part of the calculations is useful for the technical results. For legal purposes, the dispute is compared with the text of the norm and the actual conditions for its application.

The reviewer's comments are retained along with the respondent's responses. A correction, an additional procedure or a reasoned opposition may be made. The very fact of the discussion does not solve the problem. If the dispute relates to a substantive finding and is not resolved, it is referred to the competent authority or reflected as a restriction. You can't finish the review by simply removing an uncomfortable comment.

Reviewer's question	Example of defect detected
Does the object match the mission?	Different version of the model tested
Is the denominator true?	Errors are divided only to successful answers
Derogations taken into account	Complex documents deleted without explanation
Is there a conflicting material?	User complaint not included in the analysis
Is the conclusion proportional	The function test is turned into a system-wide evaluation

The review does not eliminate the underlying conflict of interest automatically. If the reviewer receives only selected positive material or depends on the same commercial decision, the organizational name of the review does not ensure its quality.

Working with external professionals

The assignment of part of the tasks to the contractor requires the identification of the subject matter, access and form of the result. An outside expert should know what question his work answers and what claims it will support. Otherwise, the report on the technical test can be used later for a legal conclusion that the specialist did not and could not make.

It is proposed to include a description of the system, criteria, permitted environment, rules for storing materials and conditions for reporting critical findings. For access to sensitive data, legal grounds and safeguards are defined separately. The contractor must not independently transfer such

data to another service for the sake of convenience of processing without prior authorisation and grounds.

The project manager remains responsible for the coherence of the outcome document. It verifies that the contractor's results are the correct version, obtained in an acceptable manner and accompanied by the necessary restrictions. In the absence of competence, the method should be evaluated by an additional reviewer rather than taking the technical complexity of the text as proof of quality.

The contractor's working materials should remain available in the agreed volume after the project is completed. The transmission of one presentation without a report may render the result unverifiable. If full disclosure is not possible, you must agree in advance on a secure view, independent confirmation or other adequate access mechanism.

For ARGA, this model allows expert work to be developed without an unconfirmed claim for its own universal laboratory. The organization can coordinate the legal question and verification of evidence by involving specialized resources within a specific task and describing their actual role.

Practical case analysis 16.07

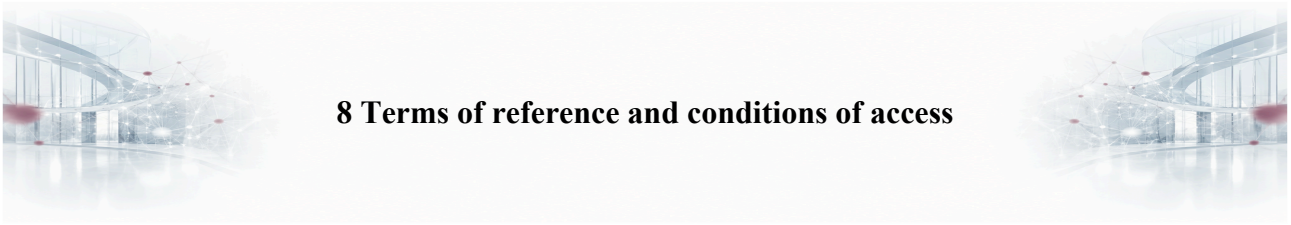
Recommendation: Designate the risk competencies of the assignment

Hypothetical example: The lawyer assesses the validity of the references, but no one checks the statistical calculation.

Analysis: Legal expertise does not compensate for an incorrect denominator or observational dependence.

Practical steps: The head attracts statistics for the plan and calculation, a lawyer for the significance of errors; appoints a reviewer.

Outcome criterion: Every significant part of the report has a competent inspector.



8 Terms of reference and conditions of access

Audited allegations instead of overall objective

The task of "checking the safety of AI" is too uncertain. It should be transformed into a list of statements: what information the system receives, what actions it performs, what errors are relevant and what measures should detect them. For each approval, the object, criterion and decision that the customer intends to take on the basis of the results of the verification shall be indicated.

For example, in a legal document processing system, you can check the correctness of references to documents, the preservation of negations, the separation of user files, and the obligation for human verification before external dispatch. These properties require different procedures. Success in one test should not compensate for the failure of another if the latter affects a substantial legal interest.

Approval	Procedure	Limit of result
Response related to source	Checking the text and the fragment indicated	Does not confirm the completeness of all sources
The files are divided.	Allowed access test and settings	Applies to tested roles
Sending requires harmonization	Checking the route and trying to bypass	Does not cover unknown integration
The error can be corrected.	Tracking up to result	Doesn't prove the same quality of all calls

The acceptance criteria shall be determined before the results are reviewed. Otherwise, the customer or the contractor can pick up a convenient indicator after problems are detected. Allowable changes to the program are documented with an explanation; the original plan and early results are preserved to show how the logic of the verification changed.

An essential question should not be left out of the task just because it is difficult to study. If a sufficient procedure is not available, it is a reason to limit the conclusion or continue the work in another way rather than replace the original question with a simpler one.

Borders of permitted access

Access must be sufficient for the selected applications and limited at the same time. The auditor does not have a common right to extract any information from the system. The task defines the environment, the records, the permitted operations, the data, the working time and the order of action in the unexpected discovery of sensitive information.

There is document access, execution monitoring, interface access and access to internal components. No level guarantees completeness automatically. Access only through the interface allows you to investigate observed behavior, but it's difficult to pinpoint all the causes of the error. Access to the code helps you understand the implementation, but it doesn't show manufacturing settings or real data.

Level of access	Possible benefits	Required reservation
Documents	Rules and declared architecture	Not yet confirmed
Observation	Real process under owner control	Could be a pre-fitting effect.
Interface	Conduct of accessible functions	Internal causes may not be known
Components and logs	Relationship of conduct to implementation	Need to check completeness and version

If the customer restricts access, the specific unavailable material and the affected conclusion are recorded. The general phrase "verification is based on the documents submitted" is insufficient when it conceals the absence of key information. The reader needs to understand the practical effect of the restriction.

For sensitive files, artificial samples and controlled copies are preferred when they actually allow the properties to be tested. Real materials are used only when there is a legitimate need and under proper conditions. The replacement of the data shall not change the subject matter of the test without being visible to the auditor.

Discontinuation and substantial finding

The contract should provide for action when a situation requires immediate attention. This can be access to someone else's documents, performing an unauthorized operation, or storing sensitive data in an external environment. Continuing the test to increase the number of confirmations could increase the damage and go beyond what is allowed.

It is suggested that the channel of notification, the person responsible and the safe stop order be determined in advance. The auditor keeps sufficient information about the find, avoiding unnecessary copying of sensitive material. A limited verification of cause and correction is then agreed. Detailed action depends on the legal regime and the nature of the system, so a universal shutdown scenario is not established.

The termination of the assignment should not destroy the trace of work already done. In the work case, the reason, the time of the stop, the procedures performed and the status of significant results are preserved. The outcome document may be a limited report on the volume achieved. It can't be considered a positive conclusion just because the whole plan hasn't been completed.

If the customer refuses to provide material material, the auditor shall assess the possibility of an alternative procedure. In her absence, the allegation remains unconfirmed. Commercial interest in completing the project does not create evidence that is not there.

For ARGA, it is suggested to include these terms in the task design before any technical inspection begins. This reduces the risk of a dispute over whether the perpetrator has the right to report a critical deficiency, to demand additional information or to refuse to make excessive use of his conclusion.



Figure 16.1. From a testable claim to a decision. Diagram of the methodology in Section 8. Evidence relates to a specific version and intended use of the system.

Practical case analysis 16.08

Recommendation: Secure Allowed Access and Stop

Hypothetical example: In order to audit the timeline, the performer has access to client letters.

Analysis: The right to change the external state is not required to read the file and increases the consequences of an error.

Practical steps: The customer limits the account to reading and stops when sensitive data is detected outside the task.

Outcome criterion: The magazine confirms that there are no unauthorized actions and that access limits are respected.

9 The audited system and the supply chain

The system is broader than a separate model

In real-world applications, the result is created by combining a model, input data, instructions, search components, access rights, and human actions. Checking only the basic model may not cover an error that occurs when extracting documents or subsequently transmitting a response. So the object is described all the way from receiving a query to getting a meaningful result.

For a legal assistant, such a card includes file uploading, text recognition, fragmentation, search, generation and user verification. There's a possible loss of meaning in every transition. For example, misrecognition of a date may lead to a correctly formulated but actually erroneous conclusion. The assessment of one last response does not always allow the cause to be localized.

Component	Audit issue
Acceptance of document	Whether the contents and the identity of the file are preserved

Component	Audit issue
Pre-treatment	Tables, denials and notes are not lost
Search for material	Which sources are available and what is excluded
Response generation	On which data and instructions the result is based
Follow-up action	Who confirms the decision and where it goes

The proposed passport of the facility lists the owners of the components and the points at which the system addresses an external supplier. This allows the defect found to be linked to the responsible party. The external component does not exempt the integrator from understanding its role in the final process.

The boundary of the check must be visible in the report. If the audit was completed on the drawing, it cannot be said that the entire decision-making process has been investigated. If a person has edited the result, the input and the possible impact of the interface on the choice is considered separately.

Identifying version and configuration

The commercial name of the system is not sufficient for re-checking. Available version identifiers, date of use, settings, system instructions and state of external sources are required. For cloud service, some information may be hidden. The auditor then saves the available identifiers and directly describes the impossibility of establishing some internal changes.

Behavior change is possible without replacing the basic model. You just need to change the query template, the tool rights, the list of documents, or the filtering rules. Therefore, the passport must cover not only the weight or the model name, but also the configuration of its use. Checks of files are useful for important components if they can be legally stored.

Passport Element	Why do you need it?
Annex version	Connect interface and integration logic
Model identifier	Clarify accessible test object
Set-up and instructions	Re-establish the conditions for receiving a reply
Composition of sources	Explain the information boundaries of the system
Rights of instruments	Identify possible actions
Date and Wednesday	Account for external changes and limitations

The check sum confirms the match of the bytes, but not the authenticity of the contents and the legality of the origin of the file. Its value is to detect a change in the material already recorded. The auditor must maintain a link between the amount, the object itself, and the way it is obtained; one list of technical values without this link is useless.

If the version is not available for re-launch, the verification can remain documentally reproducible in a limited sense: another specialist can track the calculation and conclusion on the saved materials. However, this does not amount to an independent repetition of the conduct of the original system.

Dependence on supplier

The supplier may provide a report, model description, change log or the result of his own evaluation. Such materials are useful, but their subject matter needs to be compared with verified integration. Successful evaluation of a general purpose does not necessarily include language, type of documents and the nature of errors that are significant to a particular user.

It is suggested to draw up a dependency map: which conclusion is based on one's own observations, which one is based on the supplier's data, and which one is based on assumptions. For each substantial assumption, an action is taken if it has been violated. For example, the promise of a model to remain unchanged must be accompanied by a mechanism for obtaining notifications or independently detecting a significant change.

The existence of several subcontractors makes access to evidence more difficult. The main supplier's contract may permit maintenance, but not provide the customer with sufficient information for audit. Therefore, the right to access and monitor the verification should be agreed before implementation, when another integration option is still possible.

The check must take into account the termination of the service. If the contract expires, the organization must understand which logs, settings and confirmations can be downloaded and saved. Otherwise, later investigations of the error will depend on the voluntary assistance of the former supplier. Storage time is determined by applicable obligations and necessity, not by the desire to keep everything indefinitely.

For ARGA, the principle of limited trust in supplier documents is useful: accepting them as one of the sources, checking internal consistency, and seeking direct confirmation of the most significant claims. This does not mean that a supplier is considered unfaithful in advance; it is a way to justify a conclusion regardless of its marketing description.

Practical case analysis 16.09

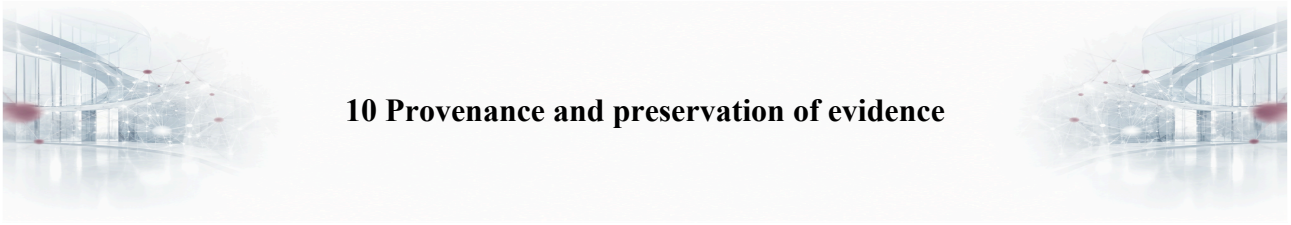
Recommendation: Describe all tested configuration

Hypothetical example: The model remained the same, but the search index was updated before the test was repeated.

Analysis: The changed result cannot be attributed only to models: the context has also changed.

Practical steps: The engineer fixes the model, index, instructions and tools versions; the comparison is repeated on the saved assemblies.

Outcome criterion: The launch card makes it possible to distinguish a model change from a search change.



10 Provenance and preservation of evidence

Inventory of materials

The working register links each material to the source, date of receipt, object and verified statement. It includes not only files, but also observations, interviews, query results and data deletion decisions. The reader should understand what material was actually used and where the saved version is.

The proposed minimum composition of the record includes an identifier, name, origin, access conditions and a brief description of the proof value. When a file is changed, a new entry or a version with a trace link is created. You can't replace the original document with an edited copy without keeping the difference between them.

Field	Example of conditional record
Identifier	E-014
Material	Test media request log
Source	Unload by the administrator in the presence of the auditor
Period	Test interval set
Question under review	Maintaining challenges to external instruments
Limitation	Completed prior to the interval

The register does not have to disclose to all participants the full path to a sensitive file. You can separate the description of the material and the secure location. The authorised reviewer shall have access to evidence when necessary to verify the conclusion.

An incomplete register creates a risk of subtle change: the new version of the supplier's report may differ from the one on which the assessment is based. Therefore, when the task is completed, the composition of the work case and the relationship between the findings and the specific versions of the materials are recorded.

Material processing chain

The material often undergoes transformations: recognition, translation, purification, depersonalization and calculation of indicators. Each conversion can change the evidentiary value. Therefore, a link between the original object and the derivative results is required, including the method of processing used and the essential decisions of the performer.

For example, when translating a legal document, the modality of the translation is checked: obligation, possibility and recommendation are not interchangeable. When recognizing the table, it is important to keep the lines and columns in line. When depersonalizing, you can't eliminate the signs that the error under investigation depends on without explaining the impact of such removal on the test.

Conversion	Risk	Monitoring
Recognition	Loss of denial or number	Comparison of significant fragments with the image
Translation	Strengthening or weakening meaning	Check by reference language specialist
Cleaning	Deletion of uncomfortable observations	Register of rules and deleted records
Impersonalization	Change of the test scenario	Assessment of the preservation of essential properties
Aggregation	Hiding differences between groups	Maintaining permissible detail

The check amount helps detect changes in the saved object after it's fixed, but doesn't show if it was fully represented originally. For downloading the log, the forming conditions, filters and time limits are also checked. The discrepancy between the number of events expected requires an explanation before the numbers are calculated.

The chain described is a working method for maintaining traceability. It does not guarantee the admissibility of the material in any court proceedings: the procedural requirements, the powers to obtain and the form of submission are determined separately by applicable law.

Discriminating information and inaccessible material

The auditor is obliged, in the sense of the chosen methodology, to seek not only confirmation of the original assumption, but also information that can refute it. For each substantive statement, it's helpful to ask the question: What observation would change the conclusion? If there is no answer, the check risks becoming a formulation of a previously accepted position.

A contradiction between documents does not always mean a breach. The policy could be updated, the report could relate to a different version, and the magazine to a different environment. The dates and the object are first specified. If the difference remains unexplained, it's kept as a finding rather than chosen as a more convenient document without explanation.

The non-availability of material shall be recorded separately from its absence. The customer may not have the right to transmit personal data or control the supplier's internal component. This circumstance affects the choice of procedure and the limitation of the conclusion, but does not in itself prove the legitimacy or wrongness of the system.

It is suggested to keep a log of requests: what was requested, why, to whom, what response was received and how it affected the program. For a substantial refusal, a possible alternative is indicated and the reason for its sufficiency or inadequacy. So the reviewer sees that the restriction has actually been investigated.

The final report describes significant contradictions close to the relevant conclusion. Moving all the restrictions to a long general list at the end of the document makes it less understandable: the user can read a positive statement without noticing that the key material is not available to him.

Practical case analysis 16.10

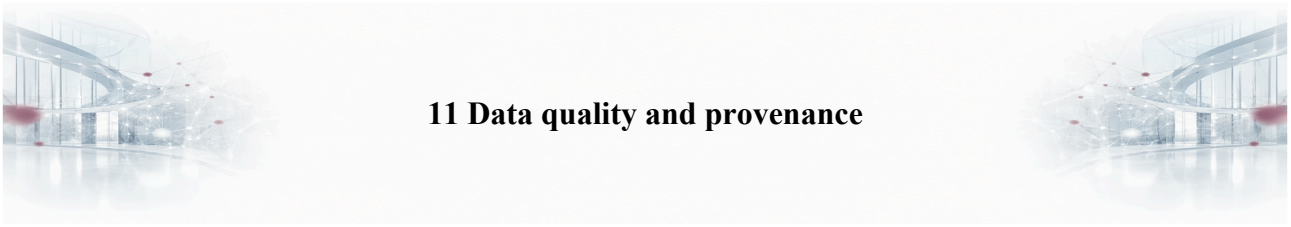
Recommendation: Save the original and the chain of change

Hypothetical example: After the scan was recognized, the table changed the dates of the two events.

Analysis: An error can be made by OCR, not by an assistant. One final text makes it impossible to distinguish between the sources of the defect.

Practical steps: The specialist stores the scan, OCR result and the corrected copy separately; indicates the conversion and verifier.

Outcome criterion: Any disputed date can be traced to the home page.



11 Data quality and provenance

The relevance of data to the question

The data are evaluated with respect to the purpose of the audit. A set suitable for displaying the interface may not be suitable for measuring errors. The archives of successfully completed cases do not represent all incoming applications. The audit must determine which set is intended to be described and why the available observations relate to it.

It is suggested to check the origin, completeness, relevance, structure of the passes and connection to the real environment. The omission itself is information: the absence of a date of birth can occur more frequently in a certain category of documents and influence false matches. Automatically deleting incomplete records can make the test artificially successful.

Data property	Verification question
Origin	The process from which the records were received
Fullness	What cases are missing and why
Relevance	Are these relevant to the period of application
Structure	Whether the links between fields and documents are preserved
Representation	Which groups and scenarios are underrepresented

If real data are scarce, specially prepared scenarios can be used. Their purpose should be clear: they help detect a specific defect, but they do not necessarily assess the frequency of that defect in operation. You can't turn the proportion of mistakes on deliberately complex examples into a total risk assessment without additional reason.

For ARGA, the properties that affect meaning are prioritized: transliteration of a name, multiple alphabets, scan quality, incomplete dates, conflicting documents, and the difference between a party's statement and an established fact. These signs are included in the data plan prior to testing.

Reference mark

Comparison of the response with the "right" result requires testing of the standard itself. In legal and social problems, the answer sometimes depends on the context or allows for several well-founded options. If the meter didn't see all the materials or used an outdated standard, the measured model error could actually be a reference error.

It is proposed to develop a labeling instruction with definitions of categories, examples and the procedure for resolving disputed cases. Some of the material is independently evaluated by two specialists. Their discrepancies are analyzed until the final calculation, as they may indicate ambiguity in the criterion. A simple vote does not always eliminate legal ambiguity.

Situation	Correct processing
A single fact	Comparison with primary material
Several permissible formulations	Evaluation of the preservation of material meaning
Legal dispute	Separate category and motivation
Lack of information	Uncertainty Mark without Fake Answer
Marker error	Correcting the master standard to preserve history

The markers should, if possible, not know which system the answer belongs to, if that could affect the evaluation. In doing so, they need to be given a context without which the content cannot be verified. A blind assessment is useful only when it does not deprive the expert of substantial data.

Once the reference is corrected, the results are recalculated according to the same rules for all the comparisons. The criterion cannot be changed solely for the response of the preferred supplier. In the work case, the original marking, the cause of change and the effect on the final indicator are preserved.

Leaking test examples and selecting convenient data

A high result may reflect the model or developer's familiarity with the test samples. If the set was used for customization, it ceases to be an independent basis for final evaluation in the original

sense. It is necessary to determine who had access to the data and what changes were made after viewing the interim results.

It is proposed to divide the data for development, refinement of the methodology and final verification. The boundary depends on the project, but it needs to be documented. After finishing the system on detected errors, it is useful to check correction on known examples and evaluate the transfer of the result to new cases separately. These two procedures answer different questions.

Another risk is the choice of convenient documents only. For example, the exclusion of low-quality scans may be justified if the system is not officially designed for them and the interface reliably rejects such files. If users regularly upload them without warning, the exclusion hides the real part of the process.

The report provides a diagram of the movement of observations: how many records were available, how many were excluded on each basis, and how many were included. For confidential data, aggregate numbers and a protected exclusion register are sufficient. It's important that the reviewer can verify that there's no arbitrary removal of adverse events.

Neither the volume nor the known origin of the kit guarantee its suitability. The main question remains: what conclusions can be drawn from this particular system and its conditions of use?

Practical case analysis 16.11

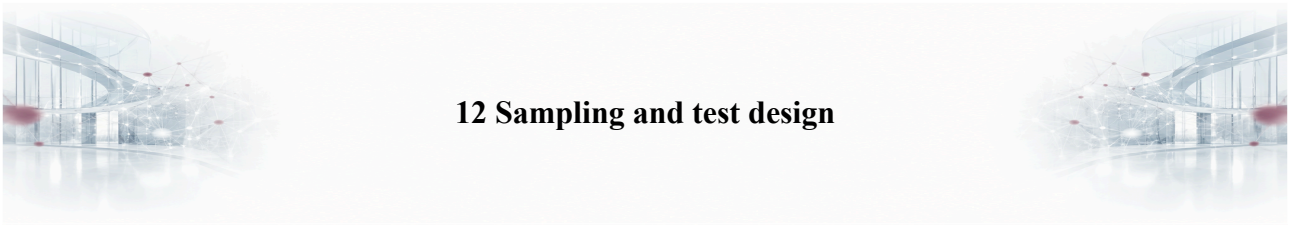
Recommendation: Checking the benchmark on mixed facts

Hypothetical example: The two experts defined the status of a client's statement differently: fact or unconfirmed message.

Analysis: The gap between the standard may look like a model error. First, we need to agree on a qualification rule.

Practical steps: The experts independently mark the material, sort out the discrepancy and check the clarified instructions on the new file.

Outcome criterion: The reference marks, the reason for the discrepancy and the version of the instruction are kept.



12 Sampling and test design

Observation unit and general population

The unit of selection in the proposed legal pilot is a file with one predefined task. Several documents inside the file, several statements in response, and repeated generations relate to the observations. The main indicator is calculated by the file, and the details of the claims are used to analyze the error mechanism. If the files belong to a single case or are built on a single template, this dependence is taken into account when selecting and evaluating uncertainty.

The selection framework consists of the permissible types of tasks, languages, volume of documents and nature of legal action. The activation, exclusion and cause of each exception shall be recorded prior to the test. The exploitative assessment requires selection that corresponds to the

claimed set; a separate set of difficult scenarios is used to find vulnerabilities. Mixing these sets is permitted to describe the work done, but not to uniquely assess the frequency of error in practice.

Decision before selection	Example for a legal assistant
Unit	One file with one question in advance
Period	Harmonized processing interval
Struts	Language, type of document and quality of recognition
Exceptions	Only pre-arranged limits on appointment
Main result	Proportion of files with significant error of meaning

If you need to infer the average error rate, random or reasoned probability selection is particularly important. If the goal is to find vulnerability, targeting complex cases can be more effective. These designs are suitable for different tasks; their results cannot be mixed into one indicator without explanation.

It is suggested that the selection rule and the original list of accessible objects be maintained. For a computer sample, the parameters used and the implementation are recorded. This allows you to check if the sample has changed after the system's answers become known.

Rare errors and test size

The absence of an error in a small sample does not prove impossible. In independent equally distributed trials with constant probability of error p, the probability of not seeing any error for n attempts is equal to (1 − p) n. This simple model helps explain why rare events require a large number of appropriate observations.

If no errors are found, the one-sided upper 95 percent limit for p in this binomial model is 1 − 0.05^1/n. For 100 tests, it's about 2.95%, for 300 - about 0.99%, for 1000 - about 0.30%. It's a calculation of uncertainty under these assumptions, not a guarantee of system security. [32]

Number of independent tests	Errors Found	Upper border near
100	0	2,95%
300	0	0,99%
1000	0	0,30%

In legal applications, even a small probability can be significant because of the severity of the consequences. Therefore, the decision should not be based on frequency alone: the type of error, the

possibility of detecting it and the mechanism for preventing harmful action are further evaluated. The absence of leakage on the test does not replace the restriction of access rights.

If observations are dependent, the same arithmetic can create false confidence. Repeating the same query a thousand times explores the variability of the answer to that query, but it doesn't cover a thousand different situations. This task requires a different plan and a separate explanation of what is measured.

Combination of sample and target scenarios

The work programme shares the operational and diagnostic sets. In the former, a well-founded set of future tasks is preserved; in the latter, the proportion of contradictions, missing documents and ambiguous identifiers is deliberately increased. The first set evaluates the quality of the field when the selection prerequisites are met. The second shows the defect mechanism and the conditions for its occurrence. For both, the original numbers, versions, and a portion of the uncertain expert score are retained.

Target scenarios are selected according to the intended route of harm. For match-finding, it can be synonyms and transliteration differences; for generalizing, it can be conflicting documents and negations; for an action tool, it can be an attempt to perform an operation without the intended confirmation. The script should verify a specific statement, not just demonstrate an unusual response.

Recruitment	Key question	The result seems to be
Representative	How often a given error arises	Undefined and transferable assessment
Target	Does the specific protection work?	Scenario, result and conditions of play
Regressive	Whether the correction was retained	Comparison before and after change
New control	Whether improvements are carried forward to new cases	Independent check after settings

The plan should take into account resource constraints. If it is not possible to verify all the languages and types of documents, the priority groups are chosen based on risk and the result is clearly limited. An honest restriction allows you to take the next step: expand the test, narrow the use, or increase human control.

The final volume is determined by the required conclusion and the nature of the risk. The universal number of tests for any AI is methodologically unfounded; it must be derived from the question, the data structure and the permissible uncertainty.

Practical case analysis 16.12

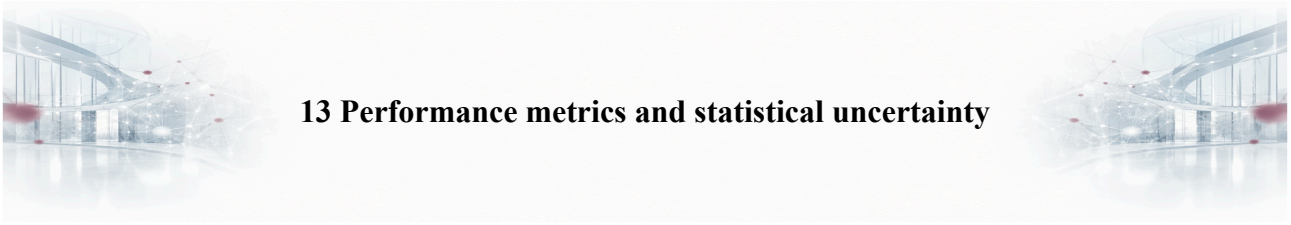
Recommendation: Segregate sample and target check

Hypothetical example: The random sample does not contain a single unreadable document, although such cases do occur.

Analysis: The average result does not show sustainability in a rare, difficult scenario. Adding targeted cases changes the composition of the set.

Practical steps: The auditor conducts a separate stress test with poor scans and reports both results separately.

Outcome criterion: The proportion of random sampling errors is not mixed with the share of stress dials.



13 Performance metrics and statistical uncertainty

Error matrix

The quality indicator should be consistent with the decision that the system helps make. The overall proportion of correct answers can hide an unacceptable number of false accusative matches or missed significant events. So for a binary result, you first make an error matrix, and then you pick the indicators that reflect the consequences of each type of error.

The hypothetical example examines 10,000 cases, of which 100 actually fall into the category sought. The system correctly found 90, missed 10, and mistakenly marked 99 other cases. A further 9,801 cases are correctly left unflagged. Of the 189 signals, 90 are true: the share of confirmed signals is about 47.62%, although the overall accuracy reaches 98.91%.

System result	Category is	No category
Positive signal	90 correct detections	99 false signals
Negative result	10 passes	9801 Correct deviation

The sensitivity here is 90/100 = 90%. The frequency of false positive results in negative cases is 99/9900 = 1%. These indicators have different denominators and answer different questions. It's dangerous to call either one of them "accuracy" without a definition: the user can understand the number as the probability of the truth of a given signal.

For ARGA, the practical investigation consists of checking each legally significant match on independent traits. A high overall percentage does not turn the signal into a known fact. The report provides the initial numbers, formulas and rounding so that the calculation can be verified without access to the author's software.

The decision threshold and the price of error

A change in the threshold can reduce false signals and at the same time increase the number of passes. In the same hypothetical set of more stringent thresholds, it gives 80 true detections, 20 passes and 20 false signals. The proportion of confirmed signals increases to 80%, but the

sensitivity decreases to 80%. The choice of the best option depends on the destination and the permissible effects.

The auditor should not choose the threshold solely according to the convenience of the final indicator. We need to figure out who's authorized to make the compromise, what data it's based on, and what safeguards compensate for the residual risk. For preliminary searching for documents and blocking operations, the same indicator can have a completely different meaning.

Indicator	Reference threshold	More stringent threshold
True Detections	90	80
Passes	10	20
False signals	99	20
Percentage of confirmed signals	47,62%	80%
Sensibility	90%	80%

The threshold is best chosen on the setup data, and finally evaluated on a separate set. If it's optimized directly from the final test, the result can be overly favorable. The history of threshold selection is preserved along with the rejected options and the motivation for the decision.

In conclusion, the unconditional word "better" should be avoided. A more accurate formulation indicates what type of error has decreased, what has increased and for what application such a change is recognized as permissible. The decision on a significant legal consequence shall be further tested under the applicable rules.

Intervals and fair presentation of results

The sample estimate is different from the unknown total. The confidence interval shows statistical uncertainty under certain assumptions, but does not automatically take into account markup errors, inconsistencies in the environment and hidden dependence of observations. Therefore, the numerical interval must be accompanied by a description of the design and significant limitations. [32]

Prior to the final selection, the main outcome, the method of assessing the difference, the rules for taking into account rejections and passes, as well as the number of basic comparisons are recorded. A substantial error and a correct refusal of an answer are coded separately: excluding refuses from the denominator can artificially improve the score. For a series of model variants, the order of comparison or adjustment of multiple checks is set in advance. An investigative analysis of additional incisions is permitted, but is indicated separately from the main examination.

When comparing configurations A and B on a file, the analysis shall retain the pairs. In addition to the total error rates, four numbers are shown: error of both configurations, error only A, error only B, absence of error in both. It's the mismatched pairs that make it possible to estimate the direction

of change; their statistical comparison is related to the McNamara approach. [M4]. If the new errors are heavier than those that have been corrected, the favorable mean difference is not a sufficient basis for admission.

Publication Element	Why do you need it?
Number of observations	Allows an assessment of the information base
Definition of error	Makes the indicator meaningful
Celebrity	Prevents mixing of different shares
Method of uncertainty	Clarify the conditions for statistical output
Exceptions and omissions	Shows the limits of actual inspection

The numbers in the report serve as educational examples and are calculated from clearly cited baseline data. For a real project, the calculation must be independently verified, and the set and code used must be kept within an acceptable amount. Arithmetic consistency is necessary, but it does not in itself confirm the suitability of all research.

Summarized configuration comparison on file set

The training example includes 200 files, each processed by configurations A and B. Significant error was detected in A in 16 files, B in 6. These findings are not enough for parity analysis: you need to know which specific files have changed the result. Let's say the next joint table.

A configuration result	B has a mistake.	B has no error
There's a mistake.	4	12
No error	2	182

In this dataset, the proportion of case files with errors fell from 8% to 3%, a reduction of 5 percentage points. Twelve initially incorrect cases were corrected, but new errors arose in two cases. In the hypothetical exact paired test, assuming independent case files, the analysis concerns 14 discordant pairs: the two-sided p-value is $2 \times (1 + 14 + 91) / 2^{14} \approx 0.0129$. This calculation applies to the stated illustrative assumptions; it is not the probability that the hypothesis is true and does not establish the system's suitability for legal use. [M4]

Before the introduction decision, two deteriorating cases are dealt with. If there's a previously missing leak or a fictitious substantive basis, the average improvement doesn't close the problem. If the files are selected manually as difficult examples, 3% characterize this set, not the actual error rate of users. The numbers and the conclusion should preserve this area of action.

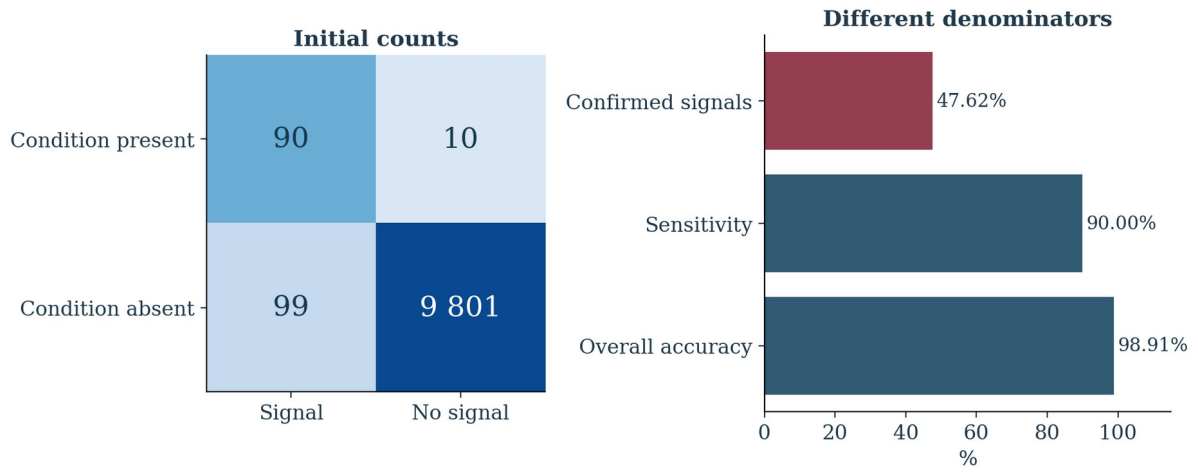


Figure 16.2. Why overall accuracy can conceal false signals. Hypothetical data from Section 13: 10,000 cases; 189 signals, of which 90 are confirmed. This is an illustrative calculation, not an ARGA finding.

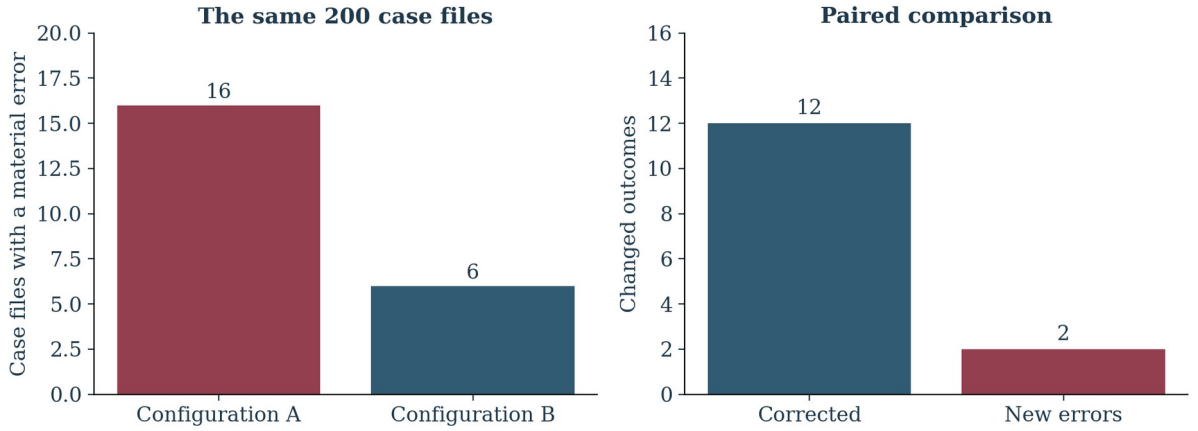


Figure 16.3. Improved outcomes and new defects. The hypothetical paired example in Section 13: the proportion of case files with errors falls from 8% to 3%. The two new defects require separate analysis.

Practical case analysis 16.13

Recommendation: Show reference quantities and new defects

Hypothetical example: For 100 identical files, version A was wrong at 10, version B was wrong at 8; version B had 3 new error files.

Analysis: The final two-fold improvement hides new flaws: five errors corrected, three introduced.

Practical steps: The analyst draws up a matching table and qualifies three new defects separately by consequences.

Outcome criterion: The decision takes into account new substantive errors, not just the average difference.

14 Bias and implications for human rights

Differences between indicators and legal conclusion

The difference in results between groups may indicate a problem, but does not automatically determine its cause or legal characterisation. At the same time, the absence of a difference in one indicator does not prove that there is no discrimination. It is necessary to identify the law concerned, the situations compared, the relevant features and how the system affects the actual decision.

NIST SP 1270 considers biases beyond one technical defect, including human and organizational sources. This is a useful reminder for the audit: problems can arise when you set a task, select data, interpret a result or device a procedure. The measurement of the model does not cover the whole social process. [23]

The proposed programme first explores the mechanism of possible harm. For example, the system may make mistakes more often in a language document, and employees may be less likely to check its negative responses. The resulting inequality arises from a combination of technical and institutional conditions. For correction, it is important to identify this combination, not just to calculate the difference.

Level of analysis	Example of the question
Problem Statement	Is the wrong sign of success used?
Data	Are the affected groups and situations represented
Model	Do mistakes differ under comparable conditions
Interface	Doesn't it point the user to a hasty conclusion?
Procedure	Is there an easy way to fix the result?

The collection of sensitive indicators for such a study requires a separate legal assessment. Good research interests do not create automatic permissions for any processing. If the necessary data are not available, the limitations of the analysis are disclosed and the permissible alternatives are discussed with a specialist.

Comparison of groups

In the hypothetical example, 100 of the 200 group A applications were selected positively, and 30 of the 100 group B applications were selected positively. The shares are 50% and 30%, the ratio of the second to the first is 0.60. This is a description of the difference observed. To find out why, you need to know more about the criteria, the comparability of the claims and the actions of the participants in the process.

The check shall take into account the intersection of the topics. The average in a large group could hide the problem in a small sub-group. At the same time, too small a division reduces the reliability of estimates and increases the risk of identifying people. Therefore, detail is determined by subject matter and legal conditions, not by the desire to get the maximum number of tables.

Group in imputed example	Total applications	Positive selection	Percentage
A	200	100	50%
B	100	30	30%

It's suggested that we keep information about the unknown group and the reasons for the absence of a sign. The removal of such observations may change the picture if the omissions are not randomly distributed. The report must specify which records are included in the comparison and how much it covers the actual set.

The correction shall not be limited to the mechanical levelling of one number. It may require changes in data, criteria, interface or review procedure. The auditor checks whether the measure actually affects the identified harm mechanism and creates a new problem for the other group.

Participation of affected persons

People who are affected by the system can report consequences that are not visible from the technical metrics. For example, a formally available complaint may require incomprehensible documents, and correcting a record may take so long that the original error continues to cause harm. Therefore, the study of the procedure should include the effective availability of protection.

The proposed methodology uses interviews, application analysis and a conditional user route. These materials complement the tests and do not replace the statistical sample. The procedure for participation should exclude pressure on applicants and unnecessary disclosure of the circumstances of their cases. The information is collected in the amount necessary for the question to be checked.

The auditor assesses the consequences for people whose appeals have not reached an official complaint. The absence of complaints may be due to a lack of knowledge about AI, a language barrier or distrust of the procedure. Therefore, a single indicator of the number of complaints should not be used as evidence of the absence of harm.

Protection property tested	Possible evidence
Clarity of notification	Checking text with intended users
Accessibility of treatment	Passing the route without expertise
Content of the review	Record of background checks
Correction of the consequences	Change of result and related entries

For ARGA, the link to the protection of rights is an autonomous value of the study. The practical recommendation should show how a person can verify a specific error, rather than limit themselves to the general promise of responsible use of technology.

Practical case analysis 16.14

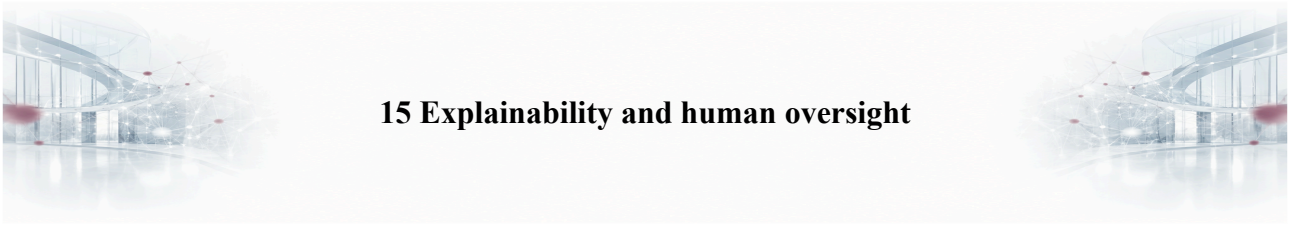
Recommendation: Check group differences on comparable tasks

Hypothetical example: For one language, short texts are used, for another, long scans; mistakes differ.

Analysis: The difference observed may be due to the quality of the entrance. It does not establish the cause or legal discrimination.

Practical steps: The researcher compares the complexity, quality of the scans and types of tasks; repeats the analysis on a comparable subset.

Outcome criterion: The report shows the composition of the groups and the persistence of the difference after comparison.



15 Explainability and human oversight

What explanation does the addressee need?

The developer, operator, regulator and affected person need different information. The technical description of architecture does not necessarily help to understand why a particular decision was unfavorable. A brief user wording may not be enough to investigate a system error. Therefore, the audit identifies the addressee of the explanation and verifies its suitability for a specific action.

In Dun & Bradstreet Austria C-203/22 the Court of Justice of the EU considered substantive information on the logic of automated processing in the context of the GDPR. It's essential to understand the procedures and principles used, not just to get a complicated mathematical formula. The application of this finding requires that the facts of the case and the corresponding right of access be taken into account. [35]

For voluntary verification, it is proposed to examine the explanation as a separate result. It must be linked to the data and procedure actually used. A plausible text generated after a decision without access to its real basis may describe a fictitious causative story. A beautiful explanation then increases the user's confidence without increasing the verifiability.

Address	Basic need
System User	Understand limitations and select action
Affected person	Establish the grounds and possibility of an objection
Technical team	Find the cause and reproduce the defect
Authorized body	Check compliance

The auditor compares the explanation with the logs, the input materials and the actual sequence. If a direct causal relationship is not established, the conclusion cannot be made that the text correctly explains the internal solution of the model.

A meaningful human test

The presence of the confirmation button does not prove human control. The employee must have the time, knowledge, access to the original materials and the authority to change the result. If the interface only shows a reliable recommendation without a basis, formal agreement can become a mechanical step.

It is suggested to investigate several conditional situations: the correct answer, a convincingly worded error, contradictions between sources and lack of sufficient data. It checks whether the employee notices the problem and what happens after refusing to agree on a response. The goal is to evaluate the actual ability of the procedure to prevent harm, not the willingness of the employee to press the right button on the instructional example.

Control condition	Verification procedure
Access to source	Monitoring of the verification of significant approval
Power to Change Result	Verification of the refusal and correction route
Enough time	Comparison of workload and complexity of the task
Competence	Disbursement without a tip from the supplier
Traceability	Record of the grounds for the final decision

Human assessment should take into account institutional incentives. If employees are rewarded only for speed, the requirement of thorough checking may not be met. The auditor reviews the instructions together with the actual load and quality control system. You can't hold a user accountable if the organization doesn't give them the conditions for substantive verification.

For ARGA, a limited initial mode is useful: the system prepares auxiliary material, and significant factual and legal claims are verified by primary sources before external use. The expansion of the authority of the system requires separate justification.

Correction and appeal

The correction procedure should not only establish the fact of the error, but also its spread. An incorrect record can be found in several documents, a search index, or a subsequent decision. The correction of one obvious answer does not eliminate the consequences. The auditor checks how the organization finds related materials and makes adjustments to the required participants.

The proposed program explores the full route: getting an application, fixing the subject, checking the original data, making a decision and confirming execution. For each step, it must be understood who is responsible. Deadlines are assessed according to the applicable law and conditions of a particular procedure; a single arbitrary period is not set for all cases.

If the decision is confirmed after a review, the user must obtain a substantive explanation within the permissible amount. Repeating the original conclusion without verifying the arguments is not

convincing evidence of effective control. For review, it is helpful to keep what new material has been reviewed and why it has or has not changed the result.

Status of treatment	What must be seen in the record
Received	Subject matter and mode of communication with the complainant
Checked	Contributions and questions raised
Resolved	Basis of confirmation or correction
Implemented	What records and actions have been changed
Closed	Communication of result and further opportunities

The audit report should distinguish between the existence of the procedure and its effectiveness. A single affirmative statement is not sufficient if the appeals under consideration show that it is impossible to correct an obvious error.

Practical case analysis 16.15

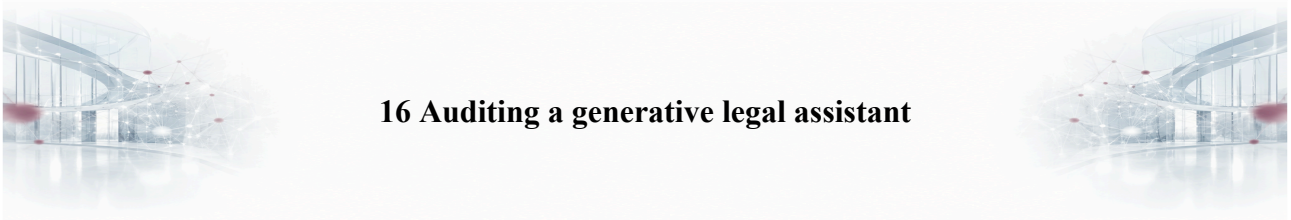
Recommendation: Check the ability to change the result

Hypothetical example: A staff member confirms the sanctions flag, seeing only a red indicator without a date of birth.

Analysis: The formal pressing of the button does not provide a meaningful identity check.

Practical steps: The process owner opens the employee's primary identification, authorizes flag revocation and verifies the educational dispute.

Outcome criterion: The staff member has, for good reason, cancelled the wrong flag; the decision is reflected in the log.



16 Auditing a generative legal assistant

Actual and legal credibility

Legal certainty in a pilot is determined by the verifiable components of the answer: actual confirmation, its documentary basis, status of certainty, applicable jurisdiction and date. An error that could alter a legal assessment or a recommended action is recognized as significant: for example, substituting a party's statement with an established fact, losing a denial, using an unapplied rule or a fabricated reference. The criterion of gravity is fixed to markings; stylistic credibility is not included.

The reference is based on the source file and the primary legal material. The two specialists independently note the binding provisions and the permissible options for the conclusion;

disagreements remain until agreed. If the experts can't justify a single qualification, the task gets ambiguous status with a description of the permissible range of answers. The model is not penalized for differing from the same wording while retaining its factual and legal meaning.

Type of defect	What the auditor checks
Non-existent source	Existence of official publication
Wrong reference	Compliance of the document with the above-mentioned question
Loss of condition	Full context of the rule and exceptions
Obsolete revision	Date and applicable version
False fact	Origin and confirmation of approval

The test uses questions with verifiable answers and pre-established significance criteria. The assessment should not encourage a reliable formulation with insufficient data. Sometimes the right behavior is to indicate uncertainty or to ask for a necessary document.

The final indicator is better accompanied by a review of critical errors. Ten stylistic inaccuracies and one false reference on which the legal position depends have different meanings. A simple averaging can hide the main risk.

Completeness of response and refusal of response

The system can increase the proportion of correct answers by answering only light questions. Therefore, it is necessary to measure the correctness of the answers given and the scope of the tasks given. In a hypothetical set of 100 questions, the system answered 60, of which 54 were correct. The correctness among the answers is 90%, the coverage is 60%, and the share of all tasks with the correct answer is 54%.

These numbers are not contradictory. They describe different properties. For an auxiliary tool, careful refusal may be useful if the user understands the reason and has a working way to continue the task. For an automatic process, a failure without a processing route can cause delay or a hidden adverse outcome.

Status in the hypothetical set	Number	Value
The Right Answer	54	Task accomplished within established limits
Wrong answer	6	Analysis of the type and effects of error required

Status in the hypothetical set	Number	Value
Failure or uncertainty	40	A separate processing route is needed

It is proposed to record which topics and groups of documents are more likely to lead to rejection. General coverage can hide systematic inappropriateness for a particular language or file type. If these cases are included in the product's claimed purpose, the restriction must be reflected in the decision on introduction.

For ARGA, an interface requirement is useful: the uncertainty must be noticeable and not formulated as a final legal conclusion. The employee needs to see exactly what is not established and what additional source will allow him to continue his work.

Source within the instrument and authority of the instrument

The material uploaded for analysis may contain text that attempts to change the behavior of the assistant, such as a request to disclose another file or perform an external action. It is important to verify that the system separates the contents of the document from the trusted instructions. Tests are conducted in a permitted, isolated environment and should not involve actual foreign materials.

The program includes safe artificial scenarios: a contradictory indication in the app, a false link, text on behalf of the administrator, and an attempt to call an unauthorized tool. It's not just the answer, it's the action that's done. The refusal of the final text does not prove the absence of intermediate data transmission.

Monitoring	Test result
Division of instructions and data	The document does not empower the system
Restrictions on access	Request not received by another person's file
Confirmation of external action	Delivery not carried out without a decision
Journaling	Significant challenges are visible and their impact

If the assistant is able to send messages or change records, the audit should cover the chain of action, not just the generation of text. Confirmation of the addressee, showing the user a specific result before sending and the possibility of stopping are of particular importance. These measures are proposals for the chosen scenario and are adapted to real risk.

Technical protection does not exempt from verification of the legal basis for processing documents. Access to the file, transmission to the supplier and storage of the logs are determined separately from successful testing.

Practical case analysis 16.16

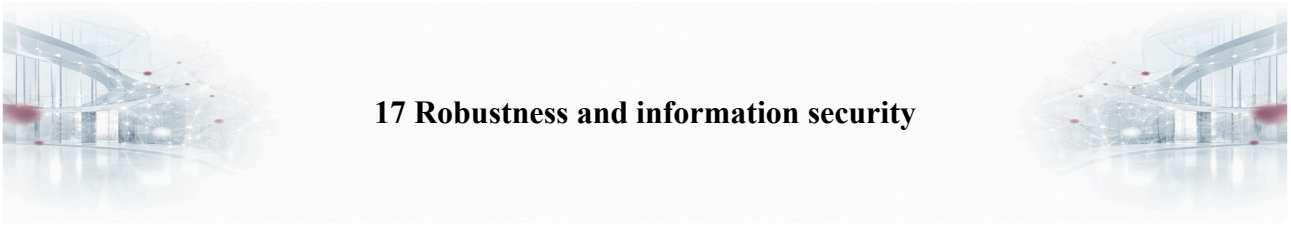
Recommendation: Check facts and completeness independently

Hypothetical example: The answer correctly quoted the decision but omitted the paragraph on exclusion from the general rule.

Analysis: The accuracy of the quote does not mean that the conclusion is applicable. The omission of the exception changes the legal position.

Practical steps: The lawyer compares the answer to a list of mandatory conditions and exceptions; the system is tested again after correction.

Outcome criterion: The answer retains the exception and indicates the relevant source.



17 Robustness and information security

Threat model for limited mission

A security audit starts with determining which assets are protected and who can influence the system. For a legal process, it's the content of the file, the user's authority, the reliability of the result and the ability to restore action. Checking without a threat model can waste resources on effective but insignificant scenarios and miss the easy path to harm.

NIST's AI 100-2e2025 publication systematizes attacks on machine learning and countermeasures, including modern risks of generative systems. It is used as a source of terminology and guidance for completeness of questions. A specific test program should take into account the violator's access and actual architecture, rather than mechanically reproducing the entire catalog. [24]

Activate	Possible impact	Subject matter
Confidential	Retrieving through incorrect rights	Isolation and permits
Inputs	Replacement or modification	Origin and integrity
System solution	Context manipulating	Sustainability of the protective procedure
External action	Use of surplus authority	Confirmation and limitation of functions
Logs	Trace removal	Access, security and completeness

It is proposed to describe the scenario through conditions, action and the expected protective outcome. This allows the verification to be repeated after correction. The general entry "security test conducted" does not show which paths were considered or left out.

Details of potentially hazardous tests are agreed with the environment owner and the relevant experts. This methodology is aimed at permitted protection inspection and does not create authority to investigate another's infrastructure without a basis.

Life cycle and supply of change

The risk depends not only on the model's behavior, but also on how it's developed, released and updated. NIST SP 800-218A examines safe development practices for generative AI and wide-purpose models. For the purpose of an audit, it is helpful to examine how the organization relates safety requirements to actual product changes. [25]

It is suggested to select a few significant changes and follow their path: task setting, code and configuration checking, testing, release authorization, and post-implementation monitoring. It is possible to determine whether the process is working in practice. The existence of general regulations does not confirm that the urgent issue has been subjected to the same checks.

Change phase	Required evidence
Request	Reason and risks
Training	Version and responsible performer
Verification	Protocol on Significant Features
Output	Decision of the authorized person
Observation	Data on post-implementation behaviour

Special attention is paid to external libraries, cloud service settings and search data. The change in such a component may affect the result without the usual release of the application. So the change map should reflect the actual chain of dependence, not just the developer's internal repository.

For critical functions, a safe option to restore or limit work is defined in advance. The auditor checks his suitability on the acceptable scenario. Having a backup doesn't necessarily mean being able to quickly restore a workable and legitimate process.

From discovery to confirmed correction

Once a defect is detected, it is necessary to separate the removal of a specific symptom from the correction of the cause. The change in the response to one known request may not solve the problem of the distinction of access. Therefore, re-testing includes the original example and new variants testing the same mechanism.

It is suggested to keep the discovery identifier, the affected version, the possible consequences, the agreed correction and the verification result after the change. If the correction changes another function, it is also examined in the required amount. Otherwise, closing one problem can create a new problem that is not reflected in the report.

Status	What does it mean
Detected	There is a reproducible or sufficient documentary basis
Analysed	Reason and field of impact clarified
Corrected by the supplier	Change claimed
Rechecked	The auditor examined the effect of the change.
Closed	Agreed criterion with specified limitations fulfilled

The disclosure of information about a defect should take into account the risk to users and the obligations in force. A public resume may describe the significance and status of correction without technical details that would cause harm. And you can't hide the unsolved problem behind the general phrase of continuous improvement.

The auditor's conclusion relates to the correction and the version studied. If the authorization, model or environment changes after the discovery is closed, the previous verification may become irrelevant. The terms of such a review shall be fixed together with the original result.

Practical case analysis 16.17

Recommendation: Confirm the correction with a repeat test

Hypothetical example: The supplier corrected the instruction after detection of a leak in the test identifier.

Analysis: Changing the settings text doesn't solve the problem and can make the useful function worse.

Practical steps: The engineer repeats the original authorized test and the independent nearby scenarios; checks for standard document processing.

Outcome criterion: The defect is not repeated in the tested set, the regular task is performed.

18 Personal data and anonymisation

Legality of processing during audit

Audit itself is a process of processing information. It is necessary to define the roles of the participants, the purpose, the legal basis, the recipients and the storage periods. It cannot be assumed that the transfer of data to the auditor is automatically allowed because the verification is aimed at compliance with the law. The content of the contract and the actual distribution of decisions should be compared. [5]

Depending on the circumstances, the executor may act as a processor, independent controller or in another design provided for by law. The name of the party in the contract does not replace the

analysis of the actual role. For a particular project, legal characterisation is carried out before access is granted, especially if the materials include sensitive information and trust documents.

The principle of minimal access is proposed: use artificial data, limited copies or protected viewing when they allow you to answer a question. If the distribution of errors cannot be verified without real data, the need is documented and accompanied by appropriate measures. Time savings are not an independent justification for the transfer of the full archive.

Decision before access	What's being recorded
Objective	What can't be solved without data
Role of the parties	Who determines the objectives and means of processing
Basis	Applicable legal and contractual condition
Interpretation	Fields, documents and recipients
Conclusion	Return, disposal or reasonable storage

Data protection impact assessment and independent auditing may use common materials, but are not interchangeable procedures. The audit carried out does not relieve the obligation to carry out the necessary impact assessment, if its conditions are laid down by applicable law.

Verification of the declaration of anonymity

The removal of a name does not automatically imply anonymity. Identification can be made possible through a combination of dates, places, occupations and external information. The verification should consider the actual context of access and the likely ways in which the information is linked. Pseudonymisation and anonymization have different effects and should not be confused in project documents.

The draft EDPB guidance 02/2026 proposes to consider anonymity with regard to the prospects of the relevant participant and the means that can reasonably be used for identification. As of the date of the report, this is a public discussion version, not a final guide. It is used to raise research issues and make proposals. [28; 31]

The proposed verification shall record the intended recipients, the additional information available to them, the technical capabilities and the timing of the evaluation. The test shall be carried out only by lawful and agreed means. The lack of successful linkage in a single limited test does not prove the impossibility of identifying all potential participants.

Question	Evaluation evidence
What signs remain	Description of fields and rare combinations
Who gets the data	Model of access and restriction of use

Question	Evaluation evidence
What information can be compared	A well-founded list of additional sources
Which was tested.	Permitted scenarios and results
When a revision is required	New data, technologies or recipients

A contractual prohibition on identification may be part of the protection, but it does not replace technical and legal analysis in itself. In conclusion, it is necessary to explain which properties of the data are established and which assumptions about the context remain relevant.

Logs and cross-border access

Queries and responses logs are often required for reproducibility, but they can contain the most sensitive part of the work. It is therefore necessary to determine what information is really needed, who has access to it, and how long it is stored. Full open-ended journaling of all files is not a neutral technical solution.

It is proposed to separate the information on the implementation of the procedure and the content of the document. Sometimes it's enough to keep the identifiers, parameters, test results and a secure link to the source material. In other cases, accurate text is needed to deal with error. The choice is based on evidentiary necessity and legal limitations, rather than a single rule to "preserve everything".

If the logs are contacted by an external supplier or a specialist from another country, the rules of international transmission and the actual recipients are examined separately. Encryption reduces certain technical risks, but it does not in itself establish the legality of transmission. The promise to store data in one region also requires verification of access and subcontractors. [5]

Material	Possible organization of access
General Protocol	Access of the project team
Sensible application	Limited number of authorized persons
Technical log	Shorter fields and unloading control
Public Summary	Deletion of personal and dangerous technical parts

At the completion of the project, the agreed procedures for storage and disposal shall be verified. The deletion of copies must not destroy the compulsory evidence, and the preservation of a work case must not become an unlimited archive of other people's data. The decision is taken on specific grounds and is reflected in the closure document.

Practical case analysis 16.18

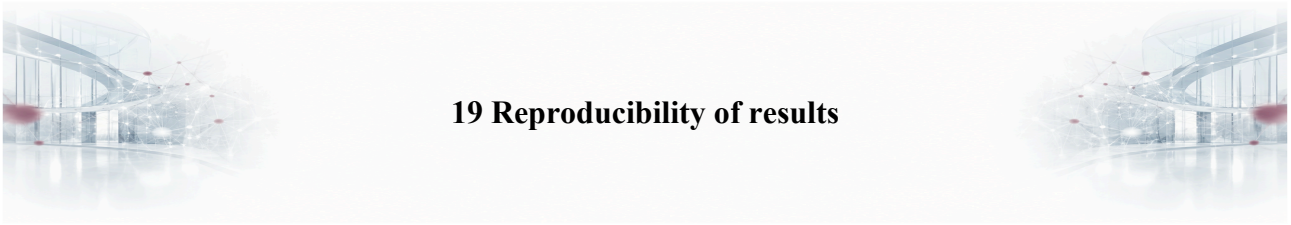
Recommendation: Minimize data without breaking the test

Hypothetical example: All names have been removed from the accountant's copy, although one name is checked.

Analysis: Impersonalization wiped out the sign of error. We need a safe equivalent with a pattern of matches.

Practical steps: The person responsible prepares the conditionally agreed identifiers and restricts access to the source separately.

Outcome criterion: The test retains one-names and different dates without revealing real persons.



19 Reproducibility of results

That another specialist should be able to repeat

Reproducibility is checked at three levels. Another researcher must restore the input data composition and configuration, repeat processing of the saved answers and get the same basic numbers. Requesting a cloud model further verifies the stability of behavior, but does not replace the reproducibility of the calculation itself. The difference between these procedures is clearly reflected in the report: the saved answer can be re-checked even when the previous model is no longer available.

It is suggested to keep the input materials in an acceptable volume, the versions of the components used, the parameters, the instructions, the sequence of actions and the unprocessed results. A final table without intermediate data does not allow you to check how rejections and errors were handled. A successful response screen does not replace a complete test report.

Level of repetition	What's being tested	Limitation
Recalculation	Formulas and final indicators	Does not confirm the receipt of the baseline data
Reprocessing	Preparation and evaluation rules	Depends on the security of the initial set
Restart	Conduct of the system in a given environment	Possible non-availability of the old version
Independent study	Sustainability of new data output	Needs comparability of assignment

The official documentation of the Inspect tool shows an example of structured record keeping. Having such a tool helps to organize records, but it does not guarantee the correctness of the tasks

and criteria themselves. The selection of the software should not replace the methodological test. [33]

For ARGA, a minimum reproducible package is useful: description of the problem, passport of the object, data register, test instructions, results and calculation. If technical repetition is not possible, another specialist should still see what the conclusion is based on and where the verified knowledge ends.

Variance of generation

The generator system may provide different answers with the same apparent request. The result is influenced by generation parameters, hidden updates, external sources and interaction sequences. So one successful launch doesn't confirm the stability of behavior, and one failed launch doesn't always allow for an estimate of error rate.

It is suggested to determine the number of repetitions for significant scenarios in advance and save all answers. You can't pick the most convincing option after a test and present it as the only outcome. If the goal is to demonstrate a specific capability, you need to name it; if sustainability evaluation, you need a different design.

It's important to consider the overall history of dialogue. The answer after a few clarifications is not going to match the first answer of another system if the task involves the same conditions. For each launch, the context available, the tools used and the user's intervention are recorded. The removal of an unsuccessful attempt and subsequent re-start should leave a record.

Variance factor	Proposed monitoring
Generator Options	Lock Available Settings
History of engagement	Save Substantive Context
External search	Specify the material and time found
Update service	Save Available version identifier
Retry	To account for all launches in accordance with the adopted regulation

Even with fixed settings, you can't promise a complete match unless the supplier guarantees proper infrastructure behavior. The correct conclusion describes the observed sustainability and the conditions for its evaluation. If there is considerable uncertainty, the decision to apply should take into account the possibility of a different answer in the future.

Automatic evaluation of responses

An automatic evaluator is considered a measuring instrument with its own area of suitability. It is tested on expertly labelled responses, including convincingly worded errors, correct rejections and ambiguous cases. Priority is given to the frequency at which a significant defect passes, not the average match of a total score. The order of responses, the length of the text and the mention of the

model name should not become hidden evaluation grounds; the corresponding effects are checked on the control pairs.

For confirmatory evaluation, the version of the evaluator, the instructions, the scale and the rules for transmitting the disputed case to the person are recorded in advance. The setting shall be carried out on a separate part of the material. The final set is not used to select a successful evaluation prompt. When a critical type of error is systematically missed, the automated assessment retains its auxiliary role, and the final classification of this type is passed on to a specialist.

Some of the results must be human verified according to a predetermined rule. All critical findings and significant discrepancies require separate treatment. If the model assessor participates in the preparation of the final opinion, its version, instructions and method of verifying its suitability are preserved in the work case.

Evaluation review	Why You Need It
Comparison with expert markings	Assess substantive suitability
Checking the same answers in different order	Show the effect of order
Safe misleading instructions	Test the sustainability of the evaluation procedure
Analysis of variances	Find categories of systematic error

Time savings are justified when the quality of the evaluation is sufficient for the chosen task and the limitations are known. If an automatic score cannot be explained or verified, it should not be the sole basis for a legally significant conclusion.

Practical case analysis 16.19

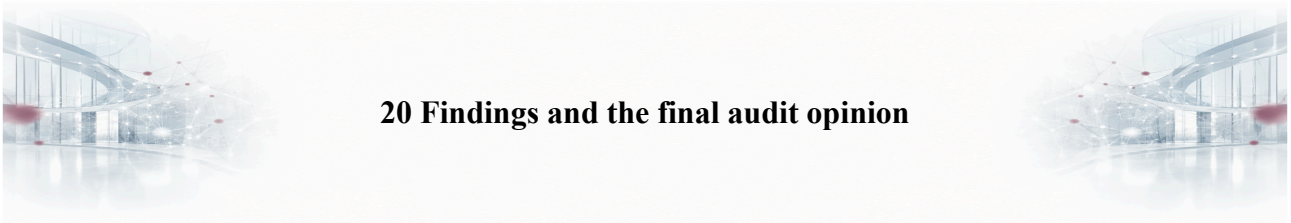
Recommendation: Transfer a reproducible startup package

Hypothetical example: The reviewer only receives the final percentage of quality without the model and mark-up rules.

Analysis: He cannot verify the calculation, the criterion, or the completeness of the registered launches.

Practical steps: The performer conveys the versions, tasks, answers, labels and calculations; the second analyst reproduces the indicator.

Outcome criterion: The independent count matches, the discrepancies are explained in the log.



20 Findings and the final audit opinion

Structure of a separate find

The finding must bind the criterion, observation and consequence. The phrase "the system is not transparent enough" does not make it possible to understand what to correct. It's more useful to

indicate that it's not possible to establish a document used for a certain class of answers, so the verifying employee can't confirm a meaningful statement until it's used externally.

It is proposed to record the identifier, object, criterion, material, observed discrepancy, consequences and necessary actions. The recommendation should be separated from the established fact. If the cause is not yet known, it cannot be presented as proven just because it seems most likely.

Find Element	A case study
Criterion	The substantial allegations relate to the original document
Observation	In 6 out of 60 responses, the fragment does not support the conclusion
Evidence	Protocol, replies and comparative documents
Value	Risk of transfer of unconfirmed allegation in the file
Proposed action	Amend reference check and repeat test

In this example, the number refers only to the set studied. Without a suitable selection, it cannot be declared as the frequency of an error of all use. If some of the answers contain different types of defects, they are described separately because the ways to correct them can be different.

The finding is being discussed with the vetted party and reviewed. A correction of the wording is permissible if it gives a more precise conclusion. The mitigation of significance for the convenience of the customer without new evidence is incompatible with the claimed independence.

Significance and priority of elimination

Significance is determined by the consequences, likelihood, extent of exposure, and detection potential prior to harmful action. It's suggested to use descriptive categories with specific criteria, not arbitrary overall scores. For a legal prohibition or leak of sensitive data, the average good performance of other functions is not a compensation.

A four-level working classification is possible: a critical problem prevents the chosen application; a substantial one requires correction or proven compensatory measures; a limited one affects the individual process; observation indicates improvement without an established significant defect. It's a project methodology, not a universal normative scale.

Category	Proposed management action
Critical	Restrict the relevant application prior to resolution of the issue
Substantive	Set the correction and test its effectiveness

Category	Proposed management action
Limited	Include in a controlled improvement plan
Observation	Consider whether a change is appropriate

The category should be explained. The same technical defect can have different meanings in the educational environment and in the preparation of a document on which a person's fate depends. The auditor takes into account the context, but should not reduce the category just because no complaint has been filed before the audit.

If the fact is established and the scale is unknown, these conditions are reflected separately. An additional scale check can be an urgent action on its own. Uncertainty is not a reason to automatically consider the problem insignificant or, on the contrary, to assert the maximum possible harm as having already occurred.

Formulation of opinion and limitations

The final opinion shall contain the object examined, the verifiable approval, the procedure, the result and the transfer limit. For each essential characteristic, one of the justified outcomes is indicated: the criterion is fulfilled; the criterion is not fulfilled; the information is insufficient; the question was not part of the task. The lack of data does not replace the established defect and at the same time does not allow positive confirmation. The conclusion is to limit the use, correct a specific mechanism, obtain missing data or continue work in that area.

The restriction must be meaningful. Instead of the general "auditor does not guarantee the absence of errors", it should be stated that no research has been conducted and how this affects the conclusion. For example, the test was performed on a single language group, there was no access to production logs or the cloud model version could not be independently recorded.

The report shall distinguish between the absence of significant identified deficiencies and the proven compliance with all declared criteria. These conclusions may require different procedures and the volume of evidence. The terms of the official scheme or standard shall apply only if the relevant procedure is actually followed and the necessary grounds are in place.

Part of the report	Mandatory content of the proposed form
Subject	Object, version, purpose and period
Basis	Criteria and their legal or treaty status
Work	Methods, data, access and significant restrictions
Result	Acknowledged and identified discrepancies
Consequences	Corrections, rechecking and conditions of use

The summary should retain the meaning of the full document. If the full report has a critical limitation, the advertisement cannot tell you that the system is fully verified and secure. The conditions under which a statement may be quoted are consistent with this logic.

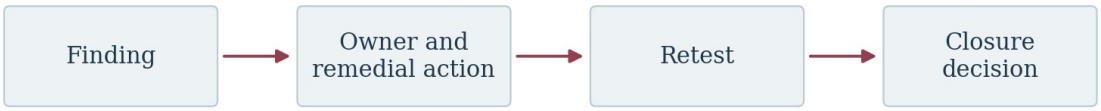


Figure 16.4. A finding and a verified correction. Each measure is linked to an owner and a retest. A claim that a defect has been corrected does not replace testing.

Practical case analysis 16.20

Recommendation: Link the find to the action to be taken

Hypothetical example: In ten responses, the reference refers to another document with a similar name.

Analysis: The general remark about poor quality does not show the mechanism and does not test the correction.

Practical steps: The auditor describes examples of replacements, assigns a document identifier check and a repeat test on nearby names.

Outcome criterion: The find has an owner, a time frame and confirmation of a repeat result.

21 Monitoring and reassessment

Audit as an assessment of a certain state

The conclusion describes the status of the system studied. After it is issued, the model, data, purpose, users or legal requirements may change. The program must therefore identify events in which the previous conclusion needs to be revised. Repeating a calendar is useful, but it doesn't replace responding to a significant change between dates.

It is proposed to include the identifier of the object, the authorized designation, the essential dependencies and the control of changes in the conditions of relevance. For example, switching from drawing to automatic sending significantly changes the consequences of an error even with an unchanged model. Such a transition should not be covered by the previous positive report.

Event	Reason for revision
New model or settings	Could be behavioral change
New language or type of document	The previous sample may not be suitable
Extension of authority	Possible harm is increasing
Change in sources	Changed information base
New legal requirement	The criteria need to be reviewed
Substantive complaint	Could be a previously unrecorded scenario.

The decision to re-check must be proportionate. Sometimes a limited test of the affected function is enough, sometimes a new program is required. It's important to document the basis, not automatically start the whole audit over or consider any change irrelevant.

For ARGA, this practice allows the analytical value of the conclusions to be maintained: the user sees not only the date of issue but also the conditions under which the results can still be applied.

Observation indicators

Monitoring should reflect the real ways of error. The general availability of the service does not indicate the validity of legal conclusions, and the average response time is the quality of human verification. Therefore, the indicators are chosen based on established risks and are associated with deviation actions.

The legal assistant may benefit from the share of unconfirmed links, loss of substantive conditions, number of corrected responses and data disclosure requests. These indicators require definitions and a reliable collection. The increase in the number of registered errors sometimes reflects an improvement in detection rather than a deterioration in the system.

Indicator	What needs to be further understood
User corrections	What changes are significant and why
System failures	What tasks and groups are affected
Complaints	Whether the channel is accessible and what is the awareness of the people
Source errors	Whether they are detected prior to external use
Change configuration	The affected properties are checked

It is suggested that quantitative data be combined with selected relevant cases. One serious incident could reveal a flaw that's missing in the big statistics. At the same time, a single unusual response should not automatically be declared a mass phenomenon without scaling.

Each signal must have a responsible and intended action: verification, application restriction, request to the supplier or re-test. An indicator system without that route creates reporting, but it doesn't provide risk management.

Incident and continuing investigative capacity

In a serious event, you need to keep enough information about the version, the input data, the actions and consequences. Unintentional cleaning of logs or immediate replacement of configuration can make it difficult to determine the cause. At the same time, the preservation of evidence should not unreasonably continue the dangerous behavior of the system.

It is proposed to determine in advance the procedure for the agreed fixation: who decides, what materials are stored, where access is restricted and how subsequent changes are noted. If the technical team is forced to repair the system urgently, the available data and its exact content are recorded before the change.

The issue of primary record	Practical Value
What happened?	Segregate observed fact from assumption
What version worked?	Link the event to a specific state
Who is affected	Assess the necessary protective measures
Which materials are retained	Ensure follow-up verification
What changed after the event	Avoid mixing the reference and corrected system

The obligation to notify depends on the applicable regime and the nature of the event. The auditor shall not automatically transfer the period from one law to another. For a specific case, the lawyer determines the grounds, addresses and deadlines, and the technical team provides the necessary facts.

The results of the investigation are being used to revise the audit programme. If the cause shows a new path to harm, it's included in future scenarios and tested for other relevant applications. The mere closure of one treatment does not always mean the elimination of systemic risk.

Practical case analysis 16.21

Recommendation: Re-rated Triggers

Hypothetical example: After the audit, the assistant was allowed to automatically create outgoing letters.

Analysis: The impact area has expanded; the previous reading test did not cover the consignment.

Practical steps: The owner suspends the new authority until the addressee has been verified, confirmed by the person and the shipment has been logged.

Outcome criterion: The new action is authorized by a separate decision on the basis of a separate test.

22 Application to ARGA's areas of work

Finding matches and working with names

When you look for matches by name, the audit task is to check how the signal turns into a conclusion about a particular person. Convergence of lines does not establish the identity of people. The value has additional characteristics, quality of the source, spelling options and the ability to correct the error.

The conditional program includes surnames, different order of names and surnames, transliteration, missing date of birth and obsolete information. Separately, false matches and omissions are investigated. You can't just evaluate a system on pre-known positive examples if in the real flow, most people aren't in the category you're looking for.

Phase	What's being tested
Collection of baseline information	Officiality, relevance and ownership of the record
Comparison	Rules for dealing with mixed features
Signal presentation	Decoupling possible overlap
Human verification	Access to additional identifiers
Corrigendum	Removal of the erroneous link and its consequences

For ARGA, a result format is useful where each match shows both confirming and contradictory signs. The final confidence of the model should not replace that explanation. If the available identifiers are insufficient, the correct condition may be uncertainty, not a forced positive or negative conclusion.

This scenario is a methodological example. It does not describe the actual architecture of INTERPOL, a bank or any particular organization, nor does it claim that a particular system has already been verified by ARGA.

Financial Analysts and Corporate Relations

In analyzing financial and corporate data, the system can detect connections between individuals, organizations, and transactions. The audit must distinguish the connection found from the verified control, violation or criminal intent. A common address, a matching name, or a sequence of translations may have several explanations.

It is suggested to check the origin of each edge of the analytical scheme: from which document it is obtained, to which date it belongs and what kind of connection it indicates. Ownership, management, representation and historical participation cannot be defined by a single symbol without explanation. An error in the communication type can alter a legal conclusion even if the name is correctly spelled.

System output	Required verification
The person is connected to the company.	Type, date and source of communication
Control in place	Applicable concept and sufficiency of facts
The operation is unusual.	Context and comparison criterion
Overstepping of the restriction	Alternative explanations and legal basis
Risk recognized as high	Value of the category and subsequent action

The verification shall take into account changes in time. The register at the current date may not confirm ownership at the time of the transaction. If the system combines different periods, the auditor explores whether an artificial chain of relationships never existed at the same time is created.

For ARGA, the practical result is a verifiable map of facts and uncertainties. An algorithmic signal may lead to further search, but the indictment requires independent evidence and legal analysis.

Synthesis and procedural protection

In the legal file, it is important to keep a distinction between the statement, the party's endorsement, the content of the document and the established circumstance. A generalization of the population can undetectably erase these boundaries. Therefore, the audit not only verifies the existence of facts in the original text but also the preservation of their probative status.

The conditional program includes contradictory explanations, denials, reservations about incomplete information and documents of different dates. It checks whether the system turns probability into reliability, lack of confirmation into refutation, or someone else's position into its own established conclusion. For such mistakes, the usual calculation of matching words is not very informative.

Reference status	Unjustified strengthening of a generalised conclusion
The Party accepts the fact	Fact established
Not submitted	Document does not exist
Communications requires verification	Control is proven

Reference status	Unjustified strengthening of a generalised conclusion
Data refer to past period	Current status
There are contradictions.	Objections excluded without justification

The evaluation shall be based on rules of reason and a subject reviewer. The system can help prepare the structure and extract the fragments, but each significant summary must remain tied to the original material. If the source is not available, this restriction is moved to the conclusion.

The original ARGA product offered is a limited verification of such an assistant on an artificial file. It will allow the procedure to be worked out without unjustified transfer of actual credentials and provide a basis for deciding whether further technical expertise is needed.

Practical case analysis 16.22

Recommendation: Maintain evidentiary status

Hypothetical example: The assistant turned the client's statement on the alleged pressure into a fact.

Analysis: The synthesis has changed the strength of the allegation and is likely to distort the treatment.

Practical steps: The lawyer enters facts, reports, and hypotheses; he checks the chronology of each source.

Outcome criterion: In the draft application, the presumption is not issued for the confirmed event.

23 Regulatory gaps and international consultations

Where existing mechanisms leave uncertainty

Comparison shows several persistent problems. Miscellaneous documents are referred to as the audit of the different subject matter. It is difficult for the customer to understand whether the result relates to an organization, model, specific application or a narrow set of indicators. A general quality mark is capable of creating a wider impression than the work actually done allows.

The second problem is limited access to supplier materials. An independent performer may have sufficient qualifications, but cannot obtain information without which a substantive statement cannot be verified. The third is the dependence on the commercial customer and the ability to control the publication. The fourth is to change the system after verification without a clear review mechanism.

Gap	Proposed course of action
Unclear subject matter of conclusion	Mandatory brief description of object and criteria
Non-transparent restrictions	Identification of inaccessible data near the conclusion

Gap	Proposed course of action
Conflict of interest	Disclosure of significant links and verification of protective measures
Non-transferable results	Passport of version, medium and permissible destination
Disruption after issuance of the report	Conditions for review and control of corrections

The proposed measures do not require any full certification audit to be announced. Their value lies in the minimum comparability of information: the user must understand what is verified and why the result deserves a certain degree of trust. More stringent requirements may be set for specific areas and consequences.

The risk of excessive formalization is also significant. A large package of documents may not be available to small organizations, and there's little to say about the actual behavior of the system. Therefore, the ARGA proposals link each requirement to a proof-of-concept and allow for proportional work without concealing restrictions.

EDPB Open Procedures

On 20 October 2026, the official pages of the EDPB indicate the deadline for discussion of the draft guidelines 03/2026 on web data collection for generative AI and 02/2026 on anonymisation: until 30 October 2026, 23:59 CET. These are two separate procedures with different topics; the proposed proposals must be addressed. [27–28]

To discuss the data collection, ARGA can offer a minimum structure of proof of origin: source, date of receipt, purpose, applicable basis, substantial limitations and history of subsequent processing. The independent verification should show which claims can be confirmed by sampling and which require more complete access. This is a proposal of the present study, not a restatement of the mandatory list of projects.

To discuss anonymisation, a position on the reproducibility of the assessment is useful: a description of the recipients, additional data, permissible linking scenarios and review conditions. The public result should retain information about significant restrictions without disclosing any material that could increase the risk of identification. The draft Guide was considered to be the subject of discussion. [31]

Procedure	Proposed ARGA contribution
EDPB 03/2026	Origin and verifiability of vendor applications
EDPB 02/2026	Evidence of anonymity and conditions for re-evaluation
General methodological proposal	Independence of the performer and disclosure of verification limits

Before submitting, the form, publication conditions and rules for the representative of the organization must be re-checked. The contribution is consistent with the established domestic procedure. Preparation of the report does not confirm ARGA's participation in the procedures; sending and receiving are recorded by individual documents. [36]

Other addresses and position priorities

The consultation of the European Commission on the impact of technology on copyright is open until 3 November 2026. It is concerned with proposals on the verifiability of the origin of teaching materials and the separation of a public resume from a confidential evidence package. The official publication directly names non-governmental organizations among the alleged participants. [29]

British ICO collects material about an IE agent until 20 November 2026. The findings of this report can be used for proposals on tooling audit, action reporting and effective human verification. This is a national procedure of the foreign regulator, which should be distinguished from EU supranational consultation. [26]

Address	Subject matter	Duration on official page
European Commission	Verification of the use of protected materials	3 November 2026
ICO	Responsibility and control of agency I.I.	20 November 2026

It is suggested to prepare short address positions, rather than sending the entire report with the same text to each body. For each position, a specific question is selected, the proposed solution and an explanation of the evidence. Additional applications are only included if they help the addressee to evaluate the offer.

Priority shall be given to EDPB procedures directly linked to independent data verification and in the near future. This is a research recommendation, not an ARGA's approved commitment to submitting materials. The actual submission, confirmation of receipt and possible publication are taken into account separately; confirmation of participation does not mean approval of the position by the regulator.

Practical case analysis 16.23

Recommendation: Formulate a targeted proposal to the regulator

Hypothetical example: The advisory response included a request to increase transparency without describing the verified record.

Analysis: The addressee does not receive a specific change in the procedure and is unable to assess feasibility.

Practical steps: The author proposes to keep the version, source and basis of the substantive conclusion; adds a hypothetical example of error recovery.

Outcome criterion: Each proposal refers to the required entry, its purpose and restriction of disclosure.



24 ARGA's practical programme and conclusions

Limited pilot project

The first pilot is proposed to dedicate the legal file to a summary, with mandatory sources and preservation of proof status. Compare two accurately recorded configurations, such as the original and the citation-checked version. The tasks for the development of the instruction are separated from the final set. The pilot checks the suitability of the error classifier, the completeness of the log and the reproducibility of the calculation; choosing the winning model is not his only goal.

Prior to launch, the research question, the main outcome and the practically significant improvement, the permissible new defects, the selection unit and the decision-maker's authority are approved. After the development, an independent test is performed; then the instruction is fixed and applied to the final set. The changes after the start of the check shall be entered in the deviation log. The final conclusion is accompanied by a verification of the original numbers and a reasoned analysis of all substantial differences.

Phase	Output
Production	Approved mission and conclusion boundaries
Training	Facilities ' passport, data and evaluation instruction
Test	Logs, initial answers and calculations
Review	Tested findings and permitted comments
Decision	Acceptable next step and modalities for revision

You can't start with a promise to confirm the reliability of the supplier as a whole. A useful initial result may be negative: for example, it is established that the system does not retain the proof status of claims. Such a conclusion allows the application to be narrowed or the procedure changed and therefore has its own value.

Implementation requires an internal decision on resources, competencies and the manner in which the outcome is to be used. The report offers a program, but does not state that ARGA has already completed the relevant tests or received official status. [36]

Minimum quality system

For sustainable expert work, it is proposed to keep a register of tasks, evidence, conflicts of interest, findings and changes in methodology. These records have to be connected. It is not necessary to create a separate bureaucratic system for each small project, but the substantive conclusion must leave a verifiable mark.

The OECD's guidelines on due diligence in the area of responsible AI are seen as an international guideline for work with adverse consequences. For ARGA, the practical sense is that research results are related to harm prevention and remediation, not just reporting. The specific programme below is the proposal of the report. [34]

Quality Element	Minimum confirmation
Competence	Designation of task specialists
Independence	Communications verification and threat decision
Traceability	Inventory of materials and procedures
Review	Verification of substantive findings by another specialist
Corrigendum	Versions of opinion and reasons for changes
Reverse	Consideration of substantiated objections

Before using the result externally, the facts, references, language and conformity of the actual work are checked. Internal approval and international agreement shall be conducted in accordance with the current ARGA procedure; this report does not replace them. Any application for accreditation, authorization or recognition requires a separate documentary basis. [36]

The criterion for the development of the programme should be considered not the number of conclusions given but the ability to reproduce a substantial conclusion and show its practical consequences. If the work doesn't allow it, increasing the volume of publications doesn't fix the methodological deficiency.

Outcome

A well-founded AI audit requires consistency in measurement, statistical inference, and application decisions. Measurement theory determines what is considered a legally significant defect; test design shows what tasks the observation covers; procedure independence allows to preserve an adverse result. The absence of any of these conditions limits the final opinion, even if the rest of the parts are done well.

For the proposed pilot, a sequence of permissions is established: first the suitability of the reference and task, then the reproducibility of the data and calculations, then the comparison of the main outcome and analysis of new defects, after which - the decision about the limits of use. A positive average does not compensate for an unrecovered critical error. An extension to another language, file type or mode of operation requires a new justification for the field of applicability.

The main risk of the market is the gap between a narrow result and a broad promise. Organizational certificates, separate tests, personnel audits and legal examinations have different subjects. They can be combined into one program, but they cannot be represented as interchangeable evidence of the general security or full legitimacy of any system.

The practical outcome for ARGA is a reproducible limited test protocol, as well as a set of verifiable requirements for an external evaluator. It is proposed to start with a single appointment of a legal assistant, confirm the functioning of the marking and re-analysis, and then decide on the extension. The methodology allows for a substantially negative outcome: it shows which mechanism needs correction and which application is not yet justified.

Further examination should determine whether this protocol helps different experts to draw comparable conclusions on the same material and identify errors essential for a legal decision. Until such data is obtained, the methodology remains an argumented research proposition. The external statement is based on the actual procedure and its outcome, not the name of the model used or the volume of the prepared text.

Practical case analysis 16.24

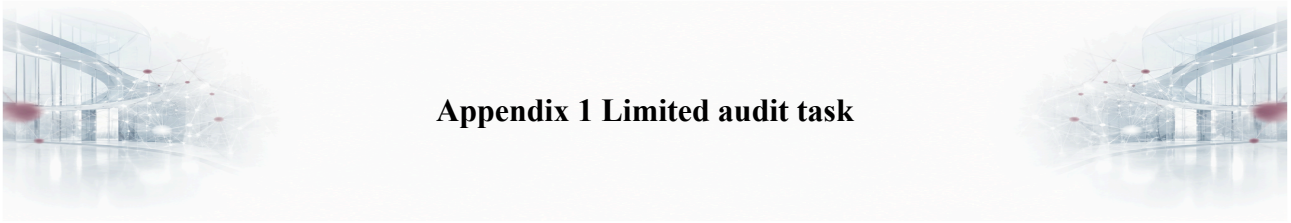
Recommendation: Start with a limited pilot and its outcome

Hypothetical example: The team is ready to check ten files, but plans to immediately declare all legal functions reliable.

Analysis: A small pilot checks the protocol's feasibility, not universal effectiveness. Without a register of errors and review, expansion is premature.

Practical steps: The curator approves a single assignment, markup, inventory and independent verification; before publishing compares the wording with the result.

Outcome criterion: It was decided to continue, finalize or stop with grounds and responsibility.



Appendix 1 Limited audit task

Completed example of subject and criteria

Below is a conditional sample of the task for the J-1.2 system, which prepares generalizations of artificial legal files. This is a draft form, not an ARGA audit record. The intended user is a lawyer who studies the sources and decides for himself how to use the text.

The D-16 test set contains 100 artificial files with one basic question to each. Including conflicting documents, denials, incomplete information and different dates. The set is specifically designed to check for significant errors; its results are not declared by an estimate of the error rate of all future applications.

Field of mission	Conditional filling
Appointment	Preparation of an outline of the synthesis with sources
Object	J-1.2 and fixed search configuration
Data	D-16, artificial documents without real files

Field of mission	Conditional filling
Main properties	Reliability, preservation of conditions, separation of materials
Deleted	External dispatch and automatic decision-making
Address	Pilot Project Manager

A reliability criterion means that a substantive statement is supported by the material and retains its probative status. If there is a lack of information, it is permissible to indicate uncertainty. A sure completion of a gap is considered a fictitious fact, even if the overall text looks convincing.

Before the work begins, the parties agree on who marks the benchmark, who executes the launch and who reviews the result. The developer can explain the system's device, but he doesn't have to determine the correctness of each contested answer alone. All changes to the task after viewing the results are subject to separate recording.

Access and conditions for completion

For the example in question, a controlled test environment, user interface, available logs and a copy of the configuration are sufficient. The supplier shall confirm the composition of the available functions. If the system turns to external services, they'll be included in the object map before testing.

Condition	Proposed entry in the assignment
Access	Own accounts with limited rights
Conservation	All initial answers, parameters and significant events
Critical find	Stopping the procedure affected and notifying the person responsible
Objections by supplier	Considered with supporting material
Review	Verification of calculations and substantive findings by another specialist
Outcome	Conclusion only on agreed appointment and version

The conclusion is not linked to the achievement of a positive result. The work is considered to be completed if the procedures have been agreed upon, substantial contradictions have been considered and a reasoned conclusion has been drawn. If the access was insufficient, the result reflects the restriction rather than hiding it under a formal task execution tag.

It is useful to have a separate verification phase for corrections in the treaty. Its volume is determined by the defects found. If the new version differs significantly, an additional launch cannot automatically be considered a continuation of the previous audit without reviewing the object and criteria.

A public description of the result is permitted only within the limits of the actual conclusion. The formula "verified the limited drafts scenario" corresponds to the subject of the example; the formula "fully certified legal AI" does not follow from the completed programme.

Appendix 2 Register of evidence

Example of related records

Each record is assigned a permanent identifier in its work history. It's used in the report and in the findings. Rename the file should not change the link to proof. If the material is updated, a history is kept that allows you to determine which version the original conclusion was based on.

Code	Material	Which confirms	Limitation
E-001	Passport J-1.2	Agreed object	Based on available service information
E-002	Recruitment D-16	Composition of tasks	Specialized examples
E-003	Expert markings	Criteria for the correct answer	Disputed cases reported separately
E-004	Background replies	Actual output of launches	Relates to the configuration
E-005	Calculation of indicators	Processing of results	Depends on the accuracy of E-003
E-006	Human Audit Log	Actions by the tester	Covers only the tasks observed

The extended record indicates the date of receipt, source, download method, access and integrity control. Confidential information about the location of materials can be kept separately from the general list. The reviewer shall have access to the required content in accordance with the procedure.

The register is useful only when a meaningful link is made to the conclusion. A hundred files without a validated statement doesn't make up a proof pack. For each substantive finding, the auditor must explain what records support it and what materials could change the assessment.

If evidence is obtained from the staff member's words, this is expressly stated. The interview is not renamed a confirmed technical fact. If possible, the explanation is compared with the log, settings, or observation in the environment.

Verification of completeness and changes

Suppose after the materials were transferred, the supplier replaced the configuration passport and reported that the previous version contained the wrong model name. The auditor saves both documents, determines the impact of the change, and checks which object was actually used in the launch. The new recording does not erase the original contradiction.

Event	Recording in a working case
New passport received	E-001-v2 with date and source
Discrepancies identified	Description of the changed field
Clarification requested	Question of the actual model used
Confirmation received	Log or other sufficient evidence
Decision taken	Save output, limit it or repeat test

If the explanation is not confirmed, the conclusion must not state an exact version that the auditor could not establish. There may be a limited conclusion as to the behaviour of the service available at that date. Its evidentiary strength is already greater than the conclusion of a fully identified configuration.

Before closing the task, the integrity of the links, the availability of key materials and the availability of the calculation are checked. It doesn't require repeating the whole study: the task is for another authorized expert to be able to recover a substantial path from the data to the conclusion.

The duration of storage shall be based on applicable obligations and agreed purposes. Once completed, the required disposal or archiving procedure shall be in place. The register does not justify the indefinite storage of personal data which is not necessary for the protection of rights and confirmation of the work performed.



Appendix 3 Conflict of interest assessment

Four conditional situations

The proposed form is used for voluntary audits and does not replace the special requirements of mandatory procedures. The decision shall be taken on the basis of the actual circumstances. The

existence of a connection does not always mean that any work is prohibited, but it can exclude the independent confirmation of a particular claim.

Situation	Significant threat	Proposed action
The perpetrator adjusted the tested algorithm	Checking your own work	Transmit the relevant evaluation to the other party
The prize depends on the positive conclusion	Financial interest	Change the condition or drop the assignment
The reviewer is linked to the vendor manager	Personal dependence	Appoint another reviewer
The buyer paid for the fixed volume of procedures	The mere fact of payment does not determine the conclusion	Check the remaining conditions and disclose the customer role

For the first situation, it is not enough to change the name of the unit if the same specialist determines the test and evaluates its result. For the second, it is not enough to declare objectivity while maintaining a dependent payment. The protective measure must affect the source of the threat.

If independent confirmation is not possible, the parties may consider consultation work with an accurate description of the role. This option does not permit the use of the word "independent" in a public restatement. The user of the result must understand that the executor was involved in creating or customizing the verified solution.

The decision on the conflict is signed by a person who is not interested in maintaining a contentious role. In the absence of such a person, you need an external competent reviewer or a refusal from the relevant part of the assignment.

Decision recording and follow-up

The conflict record should be sufficiently detailed to allow the reviewer to understand the logic. A formal "no conflict" field is permitted only after the required verification. If circumstances are unknown, they shall not be automatically considered absent.

Field	Conditional filling
The circumstances	The specialist had previously developed the J-1.2 request template
The affected task	Evaluation of errors associated with this template
Measure	Deletion of a specialist from the evaluation and appointment of criteria

Field	Conditional filling
The Allowable Role	Provision of factual explanation with fixation
Additional verification	Independent choice of scenarios and review of results
Residual limitation	The organization's overall commercial relationship is evaluated separately

The record shows that personal replacement does not necessarily eliminate the interest of the entire organization. Verification therefore continues at the level of contract, management and funding. If these links make an adverse conclusion almost impossible, the proposed measure is insufficient.

The evaluation shall be repeated when the order, payment or nature of the assistance to the employer is changed. For example, the auditor should not go from detecting a defect to developing a correction and then independently verify success without considering the conflict that arose.

The outcome document reveals significant information in the amount allowed. Confidential details of the contract may remain confidential, but this should not hide the circumstances necessary for the user to assess the independence of the conclusion.



Appendix 4 Sampling plan and rare errors

Zero error in 300 tests

The conditional task is to verify the likelihood of a specific failure of the protective measure. 300 independent tests with the same probability of error are assumed. There is no error in all attempts. In the context of this binomial model, the one-sided upper 95 percent boundary is 1 − 0,05^(1/300), that is, approximately 0,9936%.

Parameter	Value
Number of tests n	300
Errors observed	0
Unilateral level	95%
Calculation	1 − 0,05^(1/300)
Upper limit	About 0.99%

This result does not mean that the probability of error is exactly zero. Nor does it mean that the system is safe for any purpose. If a mistake can cause serious and irreversible damage, even such uncertainty can be unacceptable without further action.

With zero observed errors, the minimum number of independent trials required to obtain an upper bound no greater than 1% is determined by $0.99^n \le 0.05$. Accordingly, n must be at least 299. The calculation is valid only for the selected model and predetermined design; it does not justify a universal requirement for audit sample size.

The real project also tests the representativeness of the scenarios, the correctness of the error recording and the independence of the attempts. Without these conditions, a beautiful numerical limit can describe an artificial situation unrelated to the intended application.

When this calculation is not applicable

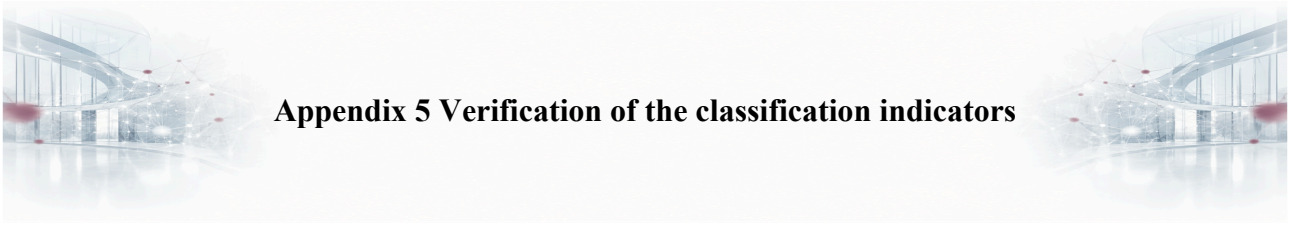
So let's say that instead of 300 independent cases, 30 documents were used, each one sent to the system ten times. Repetitions are useful for studying the variability of a response, but errors within a single document can be linked. You can't use this formula as if you got 300 independent situations without additional reasoning.

Design	What's really going on
300 independent suitable cases	Error in a given population under specified conditions
30 documents for 10 repetitions	Variance on a limited set of documents
300 specially complex requests	Resistance to selected difficult scenarios
300 known examples	Behavior on a familiar set, possible customization to test

If the check is terminated after a successful series, the stop-off rule must be taken into account. An undisclosed choice of the most favorable moment distorts the meaning of a fixed trust. The stop plan and the procedure for processing the intermediate results shall be established before the test.

The unit of error must also be checked. If a single query contains ten independent statements, the proportion of false statements and the proportion of answers differ by at least one error. For legal use, the second indicator may be more important because any significant defect can spoil the entire document.

The final entry should describe the design in words, not just the formula. The user must understand what situations are covered and why the statistical conclusion applies to them. The NIST reference is used as the basis for choosing an interval approach; the given numbers are calculated for this hypothetical example. [32]



Appendix 5 Verification of the classification indicators

Calculation of baseline quantities

Returning to the hypothetical set of 10,000 cases. A positive category is present in 100 cases. The system sends 90 true positive signals and 99 false, misses 10 positive cases and correctly rejects 9801 negative. All the figures below are derived directly from these four numbers.

Indicator	Formula	Result
Sensibility	90 / (90 + 10)	90%
Specialty	9801 / (9801 + 99)	99%
Accuracy of faithful signals	90 / (90 + 99)	47,62%
General correctness	(90 + 9801) / 10 000	98,91%
The frequency of false signals in the negative	99 / 9900	1%

The auditor checks not only arithmetic but also the meaning of the positive category. For document search, it can mean relevance; for name comparison, it can mean the identity established by the reference. It is not possible to use the same formula by changing the meaning of the category without explanation.

It is particularly useful for an outside user to show how many signals will require additional scanning. There's more false signals in this set of false signals than the right ones. This doesn't necessarily render the tool useless for a preliminary search, but it eliminates automatically accepting a signal as fact.

The calculation does not give an interval of uncertainty and does not prove the tolerability of the result. For a real conclusion, a description of the sample, the reference and the intended application environment is also required.

Change in the scope of category

Let's say the sensitivity remains 90 percent, and the false-positive rate among negative cases is 1 percent. The percentage of correct signals decreases when the required category is reduced. It's a property of calculation, not necessarily a deterioration of the algorithm.

Quantity	Correct Signal	False signals	Share of faithful
10,000 cases and 1% positive	90	99	47,62%
100,000 cases and 0.1 per cent positive	90	999	8,26%

In the second line, positive cases are also 100 and negative 99,900. One percent of the last number gives 999 false signals. So the number of positive responses a system can't interpret without knowing the context in which they arise.

For ARGA, this is particularly important in the analysis of rare matches. The supplier may report impressive sensitivity, but the user is faced mainly with the need to reject false signals. The audit task must take into account the actual flow and human verification resources required.

The premise of the constant sensitivity and frequency of false signals is here introduced for educational purposes only. In the new environment, they can also change. Therefore, the transfer of supplier indicators to another market, language or composition requires additional verification.

Appendix 6 Differences between groups

Observed ratio of results

In the conditional process, group A includes 200 applications, of which 100 were selected, group B - 100 applications, of which 30 were selected. The shares of the positive result are 0.50 and 0.30. The ratio of the second share to the first is 0.60, and the absolute difference is 20 percentage points.

Group	Statements	Selected	Percentage of selection
A	200	100	50%
B	100	30	30%
Comparison B to A	—	—	0,30 / 0,50 = 0,60

Correct description: There is a difference in the results in the test set. Wrong reinforcement: the system is proven to be committing unlawful discrimination. For a legal conclusion, the applicable rules, the comparability of situations, the mechanism of decision and other substantial circumstances are required.

One cannot ignore the distinction simply because it is not yet legally qualified. It is a basis for further verification. The auditor finds out where the discrepancy arises: in the composition of the statements, the availability of data, model rules or subsequent human actions.

The threshold of one foreign procedure is not automatically transferred to another jurisdiction. If an organization voluntarily uses a specific indicator, it is recorded as a methodological choice with justification. It should not appear to be the universal legal boundary of permissible conduct.

Additional procedures

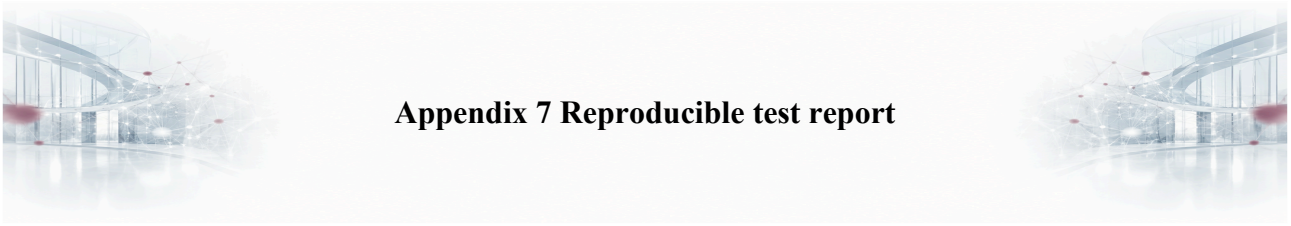
The next step is an explanation check. For example, groups may differ in type of documents or recognition quality, and the system may reject incomplete records more often. This does not eliminate the question of rights: you need to understand whether the criteria are justified and whether there are ways for people to correct the lack of information.

Possible cause	Additional verification
Different quality of source files	Comparative sub-sets and recognition checks
Missing information	Reasons for omissions and possible addition
Model features	Errors in comparable circumstances
Action by staff	Sequence of review and confirmation
Lack of access to appeal	Checking the real user route

It is recommended to check small intersecting groups when it is legal and statistically reasonable. The results with a small number of observations are accompanied by caution. You can't hide a group just for the sake of a smooth picture, but you can't make a definite quantitative conclusion based on insufficient data.

After the system changes, the comparison is repeated on independent data or on a well-founded plan. If only one indicator improves, possible new errors are investigated. The adjustment should be consistent with the problem mechanism and the legal context.

The end result of the application is a sample of the research direction, not a finished conclusion about real groups of people. All numbers are conditional and do not describe the personnel, financial or other practices of ARGA.



Appendix 7 Reproducible test report

Launch passport

The protocol is designed to preserve conditions without which repeating the result would be ambiguous. The reference example tests J-1.2 on the D-16 set. Each launch receives an identifier that is used in the evidence register and in describing the find.

Field	Conditional filling
Launch ID	RUN-016-01
Object	J-1.2, harmonized test configuration
Recruitment	D-16, set version
Instruction	Recorded text of task and system conditions
External sources	Only authorized set of documents
Tools	Search and read without external dispatch
Evaluation	Expert guidance and independent review of contentious cases
Results	All answers, refusals and errors of execution

In addition, the working protocol keeps the actual time, available model identifiers and parameters. If some settings are not disclosed by the supplier, the field is not filled in with an assumption. It indicates what information is not available and why it limits repetition.

The order of tasks can affect the outcome if the history of the interaction is maintained. It is therefore necessary to describe whether each file starts in a new context and what information is transferred between requests. Unintended mixing of the file itself can be a significant find.

The protocol shall be signed by the actual implementer in a form accepted by the organization. The manager's signature does not replace the details of who performed the procedure and what deviations from the plan occurred.

Checking a repeat result

Suppose the second expert received a different answer to the same question. First, the visible differences in terms are checked: version, parameters, dialogue history, source composition, and external search status. Only then can we discuss whether the discrepancy indicates system instability.

Comparison Step	What's being recorded
Entry check	Whether the whole material matches, not just the question.
The configuration check	Are accessible parameters and rights compatible
Context check	Is there any hidden influence of past messages?

Comparison Step	What's being recorded
Comparison of response	What significant allegations differ?
Impact assessment	Whether the user's permissible action changes

Different language does not necessarily mean a substantive defect. Two answers can maintain the same meaning and limitations. On the contrary, a slight replacement of the word "may" with "shall" could change the legal conclusion. The comparison should take into account the importance, not just the physical match.

If the cloud service update makes it impossible to run again, the original launch materials are checked. The final recording is a direct difference between re-calculation and re-start. You can't declare behavior reproducible if only the arithmetic of the final table is checked.

A register of reasons for non-reproduction is useful for subsequent projects. It shows what conditions need to be improved in the contract, logging or choice of supplier so that future conclusions are based on a more robust evidence base.



Appendix 8 Review of Legal Assistant's replies

Six significant errors in the hypothetical set

In a set of 100 questions, the system gave 60 satisfactory answers, of which 54 were accepted as correct. Six have been found to contain significant errors: three unconfirmed inferences from the source, two dates mixed and one loss of negation. In the remaining 40 cases, the system indicated a lack of information or refused to respond.

Type of result	Number	Proposed study
Correct answer	54	Monitoring the preservation of conditions and sources
Unconfirmed conclusion	3	Comparison of reference with actual content
Dating of dates	2	Audit of the provisional structure of documents
Loss of denial	1	Verification of recognition and generalization

Type of result	Number	Proposed study
Failure or uncertainty	40	Analysis of reasons and available continuation of work

These results cannot be combined into a single statement of quality 90 per cent without explanation. The number of 90 per cent relates only to the correct 60 responses. The user also needs to know the scope and the fact that six mistakes can have different consequences.

For each defect, the underlying question, material, answer and explanation of the evaluator is retained. If the gauges are separated, the case is dealt with separately. Legal ambiguity should not automatically become a system error or be excluded from analysis without trace.

In the real draft, such a table would be the result of a test. Here it is a teaching example of how to apply the methodology; no real model is evaluated by these numbers.

Decision on application and correction

For the conditional instrument, the transition to automatic preparation of final external documents is not justified. A limited auxiliary mode is possible if each essential finding is verified by a competent person at the source, and the procedure actually prevents the transmission of the error. The very promise of a human test without a proven process is not enough.

Defect	Proposed action	Recheck
Unconfirmed findings	Strengthen the link between the allegation and the source	New examples with similar but irrelevant text
Dating of dates	Individually represent periods of fact	Time-changed case file
Loss of denial	Test processing and meaningful evaluation	New instruments of denial and reservations
Excess refusals	Clarify destination and route of continuation	Comparable New Set of Challenges

Once corrected, the underlying problems should be repeated and new ones checked. The first set shows the elimination of known symptoms, the second set shows the transfer of improvement. If the supplier only demonstrates pre-corrected responses, the overall conclusion about the effectiveness of the change remains limited.

The authorized person shall decide on the permissible regime taking into account the risks and applicable obligations. The auditor provides evidence and limits to the conclusion. A positive decision should not turn a conditional pilot outcome into a promise of impeccable legal work.

If a human test does not reveal convincingly worded errors, it is necessary to change the process or limit its use. The additional instruction does not in itself prove that the problem has been resolved.

Appendix 9 Model limited conclusion

Main part of the conclusion

The quoted text is a model for a conditional draft. It does not certify actual verification and is not intended to be used as a ready-made confirmation of conformity of a particular supplier. Before applying, the form is filled in with factual information and checked according to the established procedure.

Subject of verification: system J-1.2 in configuration of preparation of black generalizations of artificial legal files. Criteria: the connection of substantive claims to sources, preservation of temporal conditions and negations, indication of lack of information. Data: specially designed set of D-16s out of 100 assignments. Real client files and automatic external shipping are excluded from the area.

The procedures performed in the hypothetical example include fixing the available configuration, starting tasks, evaluating responses and re-calculating indicators. Of the 60 satisfactory answers, 54 meet the accepted criteria, and six contain substantial errors. For 40 assignments, a refusal or an indication of uncertainty has been received. Quantities are not an estimate of all future exploitation.

Conclusion	Basis and limit
Automatic external use is not justified	Significant point defects detected
Support regime requires additional conditions	Effective verification of each relevant allegation is required
Corrections need to be checked	Change to baseline and new scenarios

The user of the result must see both successful answers and limitations. The wording of the conclusion does not make the absence of the data examined a positive confirmation. Any extension of the assignment requires a separate assessment.

Limitations and use of the document

The sample refers to the given version and specially prepared data. It doesn't explore all languages, all kinds of legal issues and the internal architecture of a cloud provider. The certification scheme, the status of the notified body or the full compliance with ISO pay standards were not assessed.

In the actual conclusion, the restrictions must be specified by the actual circumstances. If access to the logs is granted, you cannot leave a template reservation that they are not available. If part of the data has been deleted, the reasons and impact on the result should be identified. The shuttle helps the structure, but does not replace the integrity of the job description.

Completion Element	Which must be filled
Executors	Actual roles and competencies
Independence	Significant linkages and measures taken
Review	Who checked the significant findings
Conclusion version	Date, changes and current version
Use	Addresses and permissible amount of quote
Revision	Changes in which the conclusion requires updating

If a significant error is found in the report itself after it has been issued, a corrected version is released with the affected findings. You cannot silently replace a document so that the recipients of the previous edition do not know about the change that is important to their decision.

The model is not a legal waiver of responsibility for negligent work. Its purpose is to accurately identify the research done, the evidence and the limits of the use of the result. The content must be consistent with the actual actions of the perpetrators.



Appendix 10 ARGA package of proposals

Minimum quality requirements for external audit

A set of minimum requirements for disclosure is proposed for discussion with regulators and the professional community. It does not issue a single international certification and does not replace special legal procedures. The aim is to make the conclusions of different performers more clear and to reduce the risk of unfounded claims of reliability.

Proposed requirement	Evidentiary value
Name object, version and purpose	Limit the transfer of the result
Specify criteria and their status	Segregate the law from the voluntary methodology
Describe access and sample	Show actual basis of verification

Proposed requirement	Evidentiary value
Unlock substantive links	Allow me to assess independence
Save protocol and calculations	Ensure verifiability of output
Report significant restrictions	Prevent a misleading tale
Determine revision	Take into account changes in the system after audit

Proportion is achieved by the volume of procedures and the detail of the report, but not by concealing the subject. A small documentary check may have a short report if the reader sees that no technical tests have been conducted. A broader assertion requires stronger evidence.

For ARGA, this set can form the basis of an internal standard for preparing expert materials. Before adopting it, it must be adapted to the resources, competencies and types of projects. This report proposes the content of such a decision but does not state that it has already been adopted by the organisation.

Preparation of a targeted consultation position

For each external procedure, a specific problem and a proposition related to its subject matter are selected. For example, to discuss anonymisation - the reproducibility of a valuation with the data recipient in mind; to collect data - traceability of origin; to agent AI - verification of tools' powers and actions. One general text does not replace this adaptation.

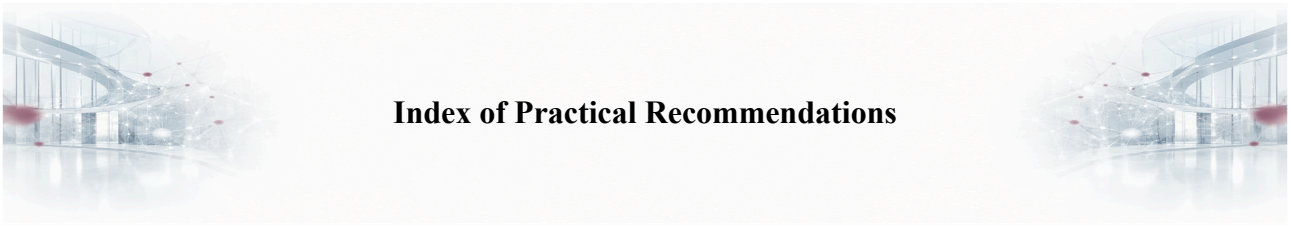
Part of entry	Proposed content
Question	What gap creates practical uncertainty
Proposal	Specific evidentiary or procedural requirement
Justification	How this helps to test and protect rights
Proportionality	How to meet the requirement without overdiscovering
Example	Artificial scenario with a clear conclusion limit
Confirmation of submission	Separate entry after actual dispatch

Deadlines confirmed for the report date: EDPB procedures 02/2026 and 03/2026 - 30 October 2026 at 23:59 CET; consultation of the European Commission on technologies and copyright - 3 November; collection of ICO materials on agent AI - 20 November. The conditions shall be rechecked prior to the direction. [26–29]

The position is internally approved according to the current ARGA procedure. For initiatives involving global governance or institutional cooperation, the international agreement provided for

shall be taken into account. The actual dispatch and receipt data are kept separate from the study itself. [36]

The final practical sequence is to select an addressee, prepare a short position, check the sources and conditions, get the necessary internal approval and only then submit the material. The presence of this report means the readiness of the analytical basis, not the participation in the consultation.



Index of Practical Recommendations

Select the recommended task and proceed with the appropriate code. The code contains the report number and the main section number. Each review provides an example, analysis, action and outcome criterion.

16.01 Fix the narrow, verifiable approval
Section 1. Outcome: The conclusion relates to the preservation of events and sources in the research set.

16.02 Map applicable requirements
Section 2. Outcome: Each requirement relates to a specific application and source.

16.03. Check the authority of the external evaluator
Section 3. Results: The contract specifies the permissible form of conclusion without extending the status of the executor.

16.04 Segregate jurisdiction and appointment of verification
Section 4. Result: For each transferable output, the basis or restriction is indicated.

16.05. Select the criterion before testing
Section 5. Outcome: Outcome contains all pre-selected indicators, including unfavourable ones.

16.06 Eliminate conflict of interest prior to appointment
Section 6. Outcome: The mechanism does not endorse the conclusion that it is effective.

16.07 Designate the risk competence of the assignment
Section 7. Outcome: Each significant part of the conclusion has a competent inspector.

16.08 Secure Allowed Access and Stop
Section 8. Results: The log confirms the absence of unauthorized activities and the compliance with access limits.

16.09. Describe all tested configuration
Section 9. Result: The start-up passport makes it possible to distinguish a model change from a search change.

16.10 Maintain the original and the change chain
Section 10. Result: Any disputed date can be traced to the reference page.

16.11 Check the reference with mixed facts
Section 11. Result: Reference marks, cause of discrepancy and version of the instruction are retained.

16.12. Segregate the sample and target verification
Section 12. Result: The proportion of random sampling errors is not mixed with the share of stress set.

16.13 Show reference quantities and new defects
Section 13. Outcome: The decision takes into account new substantive errors, not just the average difference.

16.14 Check the differences in groups on comparable tasks
Section 14. Result: The report shows the composition of the groups and the stability of the difference after comparison.

16.15. Check whether the result can be modified
Section 15. Result: The staff member, for good reason, cancels the wrong flag; the decision is reflected in the log.

16.16 Check the facts and completeness independently
Section 16. Result: The answer retains the exception and indicates the source.

16.17. Confirm correction by re-test
Section 17. Result: Defect is not reproduced in the tested set, the regular task is performed.

16.18. Minimumize the data without destroying the test
Section 18. Result: The test preserves one-names and different dates without exposing real persons.

16.19. Transfer a reproducible startup package
Section 19. Result: Independent recalculations match, discrepancies are explained in the log.

16.20 Link the find to the action to be taken
Section 20. Result: The find has an owner, a time frame and a confirmation of a repeat result.

16.21. Re-assess triggers
Section 21. Result: The new action is authorized by a separate decision based on a separate test.

16.22 Maintain evidentiary status
Section 22. Result: The draft referral does not contain an assumption for a confirmed event.

16.23 Formulation of a targeted proposal to the regulator
Section 23. The result: Each sentence specifies the required recording, its purpose and disclosure limit.

16.24 Start with a limited pilot and decisions on its outcome
Section 24. Results: A decision to continue, refine or stop has been made with the grounds and responsibilities.



References and Supporting Sources

Access date: 20 October 2026. References lead to official publications; Item 36 is an internal document. The Russian names pass on the object of the source.

[1] European Commission. EU IE Regulations, Article 113
Time frame. The portal reflects the consolidation to July 27, 2026; special deadlines cannot be postponed to all responsibilities. Official source

[2] European Commission. EU IE Regulations, Article 31
Independence and organizational requirements for notified bodies. This status is different from the usual treaty expert work. Official source

[3] European Commission. EU IE Regulations, Article 43
Different conformity assessment routes. The system category, sector legislation and applicable dates are checked for the case. Official source

[4] European Commission. EU IE Regulations, Annex VII
Evaluation of quality management system and technical documentation. Not considered to be a universal procedure for any external auditor. Official source

[5] The European Parliament and the EU Council. Regulation 2016/679 on data protection
Official publication EUR-Lex. The principles of processing, roles, safety, impact assessment and data transmission are verified on a case by case basis. Official source

[6] European Commission. Entry into force of the DSA independent audit rules
Official communication dated 22 February 2024. Annual independent audit of particularly large platforms and search engines. Official source

[7] European Commission. DSA delegated independent audit rules
Official publication on the preparation of audit reports and reports on the implementation of recommendations in the special digital services regime. Official source

[8] New York, DCWP. Automated personnel actions tools
Official description of Local Law 144: subject, bias audit, publication of information and notification. Official source

[9] New York, DCWP. Final rules on automated personnel actions
Sections 5-300-5-304. The independence of the auditor, the data, the indicators and the publication of results in the established field. Official source

[10] The British government. Introduction to confirm the reliability of I.E.
Explanation of AI contribution and differences between types of valuation. Methodological material, not universal mandatory certification. Official source

[11] DSIT. IA Independent External Evaluation Road Map
Published on September 3, 2025. Policies to develop the quality of services, competencies and the external evaluation market. Official source

[12] GAO. Artificial Intelligence Accountability Framework
GAO-21-519SP, June 30, 2021. Management, data, impact and monitoring as areas of review. Official source

[13] U.S. federal banking regulators. Revised model risk manual
April 17, 2026. Banking; generic and agent models are explicitly excluded from management. Official source

[14] NIST. AI Risk Management Framework 1.0
NIST AI 100-1, 2023. A voluntary framework; does not set a universal acceptable level of risk or a general certificate. Official source

[15] NIST. Risk profile of the generator I.I.
NIST AI 600-1, 2024. Risks of erroneous content and testing of sources in the evaluation of generic systems. Official source

[16] ISO. ISO 19011:2026
Open fourth edition record published in May 2026. Management Systems Audit Manual; full paid text has not been investigated. Official source

[17] ISO. ISO/IEC 42001:2023
Open description of the management system of the organization. The product conclusion does not automatically follow from the certificate of such a system. Official source

[18] ISO. ISO/IEC 42006:2025
Open description of requirements for the audit and certification bodies of AI management systems; addition to ISO/IEC 17021-1. Official source

[19] ISO. ISO/IEC 17029:2019
Open description of the principles and requirements for validation and verification bodies. Not used to reconstruct the full text of the standard. Official source

[20] ISO. ISO/IEC 17025:2017
Open description of the competence of testing and calibration laboratories. A specific area of competence shall be tested for application. Official source

[21] ISO. ISO/IEC 23894:2023
Open IE Risk Management Manual Card. The full list of record requirements is not restored. Official source

[22] ISO. ISO/IEC 42005:2025
Open description of the impact assessment of AI systems. Used to distinguish between Valuation Items. Official source

[23] NIST. Approach to identification and management of IA bias
NIST SP 1270, 2022. Human, organizational and technical sources of bias. Official source

[24] NIST. Taxonomy of machine training and countermeasure attacks
NIST AI 100-2e2025. Terminology and directions for threat analysis; specific tests require a solved environment and a separate task. Official source

[25] NIST. Practicals for the Safe Development of Generative I.E.
NIST SP 800-218A, final version 2024. Safe design profile, not self-certification of product safety. Official source

[26] ICO. Collection of Agent I.I. files
8 October-20 November 2026. Subjects of responsibility, transparency and data protection; terms of reference for response and publication. Official source

[27] EDPB. Advice on draft guidance 03/2026
Collection of web data for the generational AI. Official date: 30 October 2026, 23:59 CET. Official source

[28] EDPB. Advice on draft guidance 02/2026
Anonymization. Official deadline: 30 October 2026, 23:59 CET; procedure is different from discussion of web collection. Official source

[29] European Commission. Advice on technology and copyright
Publication 29 September 2026. Reception by 3 November; non-governmental organizations are among the participants. Official source

[30] EDPB. Draft manual record 03/2026
Formal confirmation of the subject matter and consultative status of the document. The proposals of the report shall not be issued for final requirements of the project. Official source

[31] EDPB. Draft guidance 02/2026 on anonymousization
Version 1.0 of 7 July 2026 for public discussion. Context of the participant and reasonably probable means of identification. Official source

[32] NIST/SEMATECH. Confidence interval for shares
Statistical Handbook, section 7.2.4.1. The numerical examples given in the report are calculated separately and are contingent. Official source

[33] AI Security Institute. Inspect documentation for test logs
Official description of the retention of the evaluation results. The programme tool does not guarantee the validity of the criterion or the independence of the evaluation. Official source

[34] OECD. Due diligence guide for responsible AI
Publication 2026. The international guideline for work with adverse consequences used within the scope of the study. Official source

[35] The EU Court of Justice. Dun & Bradstreet Austria, C-203/22
Decision of 27 February 2025, ECLI:EU:C:2025:117. Contental information on processing logic in the context of the right of access. Official source

[36] ARGA Observatory. Analytical and international institutional work plan
Internal plan of 8 October 2026. Theme No. 16, criteria of the report and procedure for agreeing external positions; not external normative source.

Scientific and Methodological References

[M1] Chouldechova A. et al. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. 2024. Scientific pre-print. https://arxiv.org/abs/2412.01934

[M2] Raji I. D. et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. FAT* 2020, pp. 33–44. https://doi.org/10.1145/3351095.3372873

[M3] Keller A. et al. Expanding the AI Evaluation Toolbox with Statistical Models. NIST AI 800-3. 2026. https://doi.org/10.6028/NIST.AI.800-3

[M4] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 1947, pp. 153–157. https://doi.org/10.1007/BF02295996