Certification Is Not Evidence

Evidence Classes and the Assessment of Safety Arguments in Battery Energy Storage Procurement

Eduardo Mayer Fagundes, M.Sc.
Electrical Engineer — Energy and Intelligent Systems
nMentors Engenharia
São Paulo, Brazil
eduardo.mayer@nMentors.com.br

Abstract — On 13 August 2026, Clearstone Energy's appeal for a 150 MW storage system in Axminster, England, was dismissed because the inspector was not persuaded that deflagration mitigation strategies would work effectively for cabinets spaced 10 cm apart. Separation distance is among the parameters that UL 9540A installation-scale testing allows to be assessed. The refusal did not follow from any breach of standards: it followed from available evidence failing to cover the deployed configuration. A previous paper by this author argued that the capacity to sustain a safety argument is a property of the control architecture, and translated that argument into technical specification items. This paper addresses the next question, which Brazil's LRCAP 2026 capacity auctions for battery storage — and the technical qualification that precedes them — place before developers, insurers, lenders and licensing authorities: once the vendor has answered, how does one assess whether the answer demonstrates anything? The central proposition is that **there is no strong evidence in the abstract — there is evidence adequate or inadequate to a given claim, configuration and operating regime**. Three theses follow. Epistemically, each claim admits only certain classes of evidence, and most procurement disputes arise from accepting a class below what the claim requires. Structurally, the absence of a higher-level test is in itself neither favourable nor unfavourable: its meaning depends on why testing stopped where it did. Temporally, evidence has a scope of validity bounded by the edition of the method, the identity of the tested configuration, and the presumed operating regime. The paper proposes seven evidence classes ordered by recoverable information and differentiated, at the upper end, along two independent axes; an explicit sufficiency criterion; a correspondence rule between claim and minimum required class; a claim class concerning authority arbitration among controllers; and a taxonomy of five modes of non-response. It applies these instruments to a constructed dossier and states explicitly what it does not demonstrate.

Index Terms — *battery energy storage systems, safety case, evidence classes, UL 9540A, NFPA 855, authority arbitration, technical specification, technical due diligence.*

I. Introduction

Brazil's Energy Research Company (EPE) closed registration for the Capacity Reserve Auctions dedicated to battery energy storage with 6,091 projects totaling 296.8 GW, for tenders scheduled for December 2 and 4, 2026 that will contract only a fraction of that volume [1]. Those auctions — designated LRCAP 2026 — are the first in the country reserved exclusively for storage, with supply obligations beginning in August 2028 [2]. Between registration and the auction lies the technical qualification stage conducted by EPE. Between qualification and operation lie environmental licensing, supply contracting, insurance underwriting, and the structuring of project finance. At each of those gates, someone has to read a set of documents and decide whether it demonstrates what it claims.

A previous paper by this author argued that compliance achieved and a demonstrable safety argument are different things, that only the second sustains a license, financing, and an insurance policy, and that the capacity to sustain one is a property of the control architecture rather than of the equipment [3]. That paper closed with specification items and with an observation left undeveloped: a certificate transfers to the buyer the conclusion of a test, not the competence to assess its scope.

This paper develops that observation. The question is no longer what to require. It is what to do with what comes back.

The instruments proposed here are not jurisdiction-specific: evidence classes, a sufficiency criterion, and a taxonomy of non-response apply wherever a buyer, an insurer, or a licensing authority must judge evidence produced by someone else. What is specific to Brazil is the calendar, and the fact that the country will contract a fleet of gigawatt scale before it has any failure case record of its own.

The case that motivated the previous paper admits a re-reading from this angle, and the re-reading is instructive. Clearstone Energy submitted the Axminster Energy Hub project to East Devon District Council in early 2024, under application 24/0096/MFUL, having secured a connection at National Grid's Axminster substation. The council refused it in May 2025. During the appeal the council itself withdrew its objection, at a meeting on January 14, 2026, leaving the opposition to local groups admitted as Rule 6 Parties — the status that gives third parties a formal role in a United Kingdom planning appeal. On August 13, 2026, planning inspector Helen Heward dismissed the appeal in a seventeen-page decision, even though the developer used equipment from a leading manufacturer and had followed the applicable fire-safety guidance.

The grounds are what matter. The inspector recorded that she was not persuaded there was sufficient robust evidence to demonstrate that deflagration and explosion mitigation strategies would work effectively, or that explosion risks would be satisfactorily controlled, for cabinets spaced 10 cm apart. Because the pollution-control evidence was predicated on the fire-safety evidence, the insufficiency of the first dragged down the second.

Separation distance between units is among the parameters that UL 9540A installation-scale testing allows to be assessed. Section V shows why the evidence typically available does not cover that parameter, and why the gap is structural to the method rather than the failing of any particular vendor.

There is a nearby precedent that exposes the question in its purest form. In the appeal concerning the Pound Road development in Hawkchurch, the fire and rescue service raised no objection, and the inspectorate refused permission all the same, recording that it had not been demonstrated that the proposal would not pose a significant risk to residents and to the environment. The technically competent authority did not object; the evidentiary authority was not persuaded. **Absence of objection is not presence of evidence.**

The central proposition of this paper is that there is no strong evidence in the abstract. There is evidence adequate or inadequate to a given claim, under a given configuration and a given operating regime. From that follow the three theses stated in the abstract — epistemic, structural, and temporal — and the structure of the text.

Section II situates the work in the safety-argumentation literature. Section III characterizes the asymmetry of evidentiary competence. Section IV proposes the evidence classes and the sufficiency criterion. Section V develops the UL 9540A case. Section VI establishes the correspondence rule. Section VII adds a claim class absent from the previous paper: authority arbitration among controllers. Section VIII presents the taxonomy of non-response. Section IX addresses the scope of validity. Section X applies the instruments to a constructed dossier. Section XI discusses institutional use. Section XII states what this analysis does not demonstrate.

II. Related Work

The idea that safety is demonstrated by a structured argument, and not by a compliance checklist, is not new. It consolidated in the United Kingdom after the Piper Alpha disaster and the Cullen report, migrating into nuclear, rail, and defense regulation — notably through Def Stan 00-56, which replaced prescription with a requirement for a safety argument [4]. Kelly and McDermid formalized the representation of that argument in Goal Structuring Notation (GSN) [5], subsequently maintained as a community standard [6], and the parallel Claim–Argument–Evidence tradition established an equivalent vocabulary. ISO/IEC 15026 standardized the *assurance case* structure for systems and software [7], and ISO 26262 brought the safety argument into the automotive lifecycle.

That framework has already been extended to systems with learned components. The AMLAS methodology organizes the production of assurance evidence for machine learning in autonomous systems [8]; UL 4600 explicitly adopts a safety-argument-based approach, rather than prescriptive requirements, for autonomous products [9]; and ISO/IEC TR 5469 addresses the properties, risk factors, and methods of using artificial intelligence in safety-related functions, of using non-AI functions to safeguard AI-controlled equipment, and of using AI in the development of safety functions [10]. In the storage domain, the application of system-theoretic process analysis (STPA) proposed by Sandia treats battery-management-system failures, flammable gases, arc flash, fire, and explosion as a problem of inadequate control rather than of isolated component failure [11].

There is also a body of work on the **sufficiency** of the argument, and not only on its form. Bloomfield and Rushby discuss confidence in *assurance cases* [12]; Goodenough, Weinstock, and Klein propose eliminative argumentation, in which the strength of the argument derives from the systematic elimination of *defeaters* — reasons why the conclusion might not hold [13]. That line is the closest antecedent to the taxonomy proposed in Section VIII of this paper: the substitution modes described there can be read as defeaters left uneliminated, and the debt is acknowledged.

This work differs from the above in four respects, which delimit its contribution.

Perspective. GSN, CAE, AMLAS, and UL 4600 are producer instruments: whoever designs the system builds the argument and submits it. The problem addressed here is the **evaluator's** — licensing authority, insurer, lender, or buyer — who must judge an argument they did not build, about a system they did not design, with asymmetric competence and against a counterparty's adverse commercial incentive.

Access. Assurance-case methodologies presuppose access to development artifacts. In procurement, the evaluator has only what the vendor discloses, typically in a format that discards the premises — hence the centrality, in this paper, of the distinction between certificate and report.

Granularity of evidence. In the existing notations, evidence is a terminal node of the argument, treated as given. This paper proposes classifying it by recoverable information and by two axes of applicability — configuration fidelity and condition realism — and shows, in the UL 9540A case, that one and the same document can be favorable evidence as to one level and an absence of evidence as to another.

Temporal window. The assurance-case literature deals with the system lifecycle. Procurement has a deadline: the contractual leverage that makes evidence obtainable expires at signature, which turns the classification of evidence into a specification decision rather than an audit one.

III. The Asymmetry of Evidentiary Competence

The market for procuring storage systems operates under asymmetries of several kinds: of information, of access to evidence, and of incentives. This paper addresses another one, which survives all of them: **even when the information exists and is disclosed, an asymmetry of evidentiary competence remains.** Assessing the delivered document requires knowledge the buyer typically does not have and that the document itself does not transfer.

A certificate is the canonical example. It communicates that an object was submitted to a procedure and obtained a conforming result. It generally identifies the model or family, the edition of the standard, and the laboratory. It does not communicate, with sufficient granularity, under what conditions the test took place, against what criteria, with what detailed configuration, or what remained outside the scope. That information exists — it is in the test report, which does not circulate with the commercial proposal. The certificate is the conclusion detached from its premises.

The buyer's reflexive response is to demand more certificates. This aggravates the problem rather than solving it, for two reasons. First, because it multiplies conclusions without recovering premises, increasing documentary volume without increasing the capacity to judge. Second, because it shifts the decision criterion to a metric — the number of certifications — that is orthogonal to the relevant question and that vendors optimize with ease.

The distinction between a product standard and a test method illustrates the confusion well. UL 9540 is the product safety standard for storage systems and their equipment. UL 9540A is not a certification standard: it is a standardized test method that generates data on thermal propagation behavior under extreme conditions [14]. Saying that a system "has UL 9540A" is categorically equivalent to saying that a sample "has chromatography." The method was applied; the result is another matter, and it is the result that answers the question.

This asymmetry is not resolved with more documents. It is resolved with a reading criterion — which requires, at a minimum, a typology of evidence classes and a rule associating each claim with the minimum class capable of sustaining it.

IV. Evidence Classes and the Sufficiency Criterion

Table I — Evidence classes.

Class	Description	What it authorizes you to conclude
E0	Vendor statement: brochure, questionnaire response, commercial presentation	That the vendor asserts it. Nothing about the system.
E1a	Organizational or management-system certification	That the organization maintains documented processes. Nothing about the product offered.
E1b	Product or subsystem certification	That an object was assessed against product requirements and obtained a conforming result. Scope only partially recoverable: conditions, detailed configuration, and results remain in the report.
E2	Test report without declared level, edition, and configuration	That measurement occurred. Transferability to the delivered object is indeterminate.
E3	Test report with declared level, edition, configuration, and criteria	That the identified object, under the identified conditions, exhibited the measured behavior.
E4	Closed-loop test with the real component under design conditions: C-HIL; worst-case execution time under thermal throttling	That the component to be delivered, with its firmware and its thermal envelope, behaves as measured under the specified condition.
E5	Instrumented operating record or forensic investigation of an incident in a comparable asset	That the behavior occurred in real operation, under conditions the designer did not anticipate.

A. The Scale Is Ordinal Only up to E3

From E0 to E3 the progression is strictly ordinal, because each class recovers information that the previous one discarded about the same object: from assertion to assessment, from assessment to data, from data to data with scope. No claim is better sustained by E1b than by E3 of the same object.

Above E3 the ordering ceases to hold, and this needs saying plainly, so as not to invite the reading that a higher class is always better. E4 and E5 differ along two independent axes:

- **Configuration fidelity** — proximity between the observed object and the object to be delivered: chemistry, module configuration, PCS, BMS, firmware version, enclosure, suppression system.
- **Condition realism** — proximity between the observation condition and the unanticipated adverse operating condition.

E4 is strong on the first axis and weak on the second: it measures the component that will be delivered, under conditions someone had to imagine. E5 is the inverse: it

records what actually happened, but in another asset, with another chemistry, another firmware, another climate, and another era.

A forensic report on a 2019 NMC system in a temperate climate is valuable input for identifying failure modes, and it is not evidence superior to a C-HIL test specific to the LFP system that will be delivered in 2028. **The classes do not establish absolute dominance over one another; final evidentiary strength depends on the correspondence between the evidence and the claim.** A dossier presenting only E4 and one presenting only E5 have different gaps, not different degrees of the same gap.

B. The Sufficiency Criterion

Class is a necessary and not a sufficient condition. The criterion is stated in predicative form:

> Evidence *e* sustains a claim *a* if, and only if: (i) the class of *e* meets the minimum class required by *a*; **and** (ii) the configuration tested in *e* corresponds to the configuration offered; **and** (iii) the operating regime presupposed in *e* covers the contracted regime.

Two consequences deserve explicit statement, because they are frequently violated in practice: **a class equal to or above the minimum does not imply sufficiency.** E4 evidence obtained on a different firmware version does not satisfy the claim, however high its class. An E3 report on a different module arrangement does not satisfy it either. Condition (i) is the easiest to verify and the least decisive; conditions (ii) and (iii) are where claims actually break, and they are treated in Section IX.

C. Three Additional Properties

Availability does not track strength. E5 depends, by construction, on something having happened. As argued in the previous paper, Brazil has no forensic case record of its own in grid storage, which makes E5 accessible only by importing foreign reports — precisely the class whose configuration fidelity is lowest [3].

The strength of a set is limited by the weakest link in the chain of support. At Axminster, the pollution-control evidence was predicated on the fire-safety evidence; the insufficiency of the second invalidated the first regardless of its own quality. Evidentiary chains inherit the strength of their antecedent.

The jump from E1b to E3 normally requires no new test. The report exists; the certificate derives from it. Obtaining it may require a confidentiality agreement or access control, and in some cases it runs into laboratory restrictions or intellectual-property protection — but it recovers evidence already produced, without imposing a new experimental campaign. It is the intervention with the best ratio of cost to evidentiary gain in the entire procurement process.

V. Anatomy of a Test Method: UL 9540A and the Stopping Logic

UL 9540A is the only consensus test method explicitly cited by NFPA 855 for large-scale fire testing, and the method specified by both the 2026 edition of NFPA 855 and the 2024 edition of the International Fire Code [14], [15]. It progresses through four levels, each feeding the next with measured parameters.

Table II — UL 9540A levels and what each one demonstrates.

Level	Object	What passing demonstrates	What remains outside
Cell	Individual cell forced into thermal runaway in a pressurized vessel	Runaway characteristics; gas composition and flammability; lower flammability limit, peak pressure, and burning velocity	Cell-to-cell interaction
Module	Module in production configuration	Cell-to-cell propagation tendency; heat and gas release rate; ignition or deflagration potential	Module-to-module propagation and unit behavior
Unit	Complete system, with suppression disabled	Module-to-module propagation; unit release rates; deflagration and reignition potential	Unit-to-unit propagation; separation distances; effectiveness of installed suppression
Installation	Deployed configuration, with active suppression	That runaway in one unit does not propagate to adjacent units in the tested arrangement; evidence for assessing separation distances and active protection	Conditions outside the tested scenario

The property that matters is the **stopping logic**: testing begins at the lowest applicable level and does not proceed when the performance criteria are met [16]. If the module test passes, the unit and installation tests are not required.

From this follows the inversion this paper wants to make explicit. **A dossier containing only cell and module tests reports a good result — the system stopped there because it passed. And, for exactly that reason, it says nothing about unit-to-unit propagation, about separation distances, or about the effectiveness of suppression in the deployed arrangement.** The same document is

simultaneously favorable evidence as to module behavior and an absence of evidence as to the installation.

The naive reading is that more levels tested means a safer system. The correct reading is that more levels tested means the earlier levels exhibited behavior that required proceeding — and that, at the same time, information exists about the scale at which the asset will actually be installed.

This gap is not this author's interpretation. It was recognized in the codes and corrected. The published technical justification for the 2026 revisions is explicit: because the method can end early when a level passes, some large systems were never tested as a whole [16]. The 2026 edition of NFPA 855 began requiring large-scale fire testing in conjunction with UL 9540A, introducing Annex G.11, which gives guidance on initiating a developed fire condition, on assessing separation distances, and on the role of active protection systems [15]. The 6th edition of UL 9540A, published on March 13, 2026, revised Section 10 to incorporate a large-scale test method aligned with Annex G.11, dropped the unit-level test for non-residential systems — which go from module level directly to installation level — and added Annex C, for large-scale deflagration testing of enclosures [14].

Three consequences for anyone reading evidence in Brazil.

The first is temporal. Reports issued under earlier editions were produced under a method that permitted precisely the gap the 2026 edition closed. A 2024 dossier can be technically impeccable and still silent as to installation scale, and the report date is the indicator of that condition.

The second is one of geographic scope. The large-scale testing requirement derives from North American codes. The Brazilian framework has no equivalent: no domestic code obliges a developer to test at installation scale, and no Brazilian authority will request that report on its own initiative. The gap NFPA 855 closed remains open here, and the Brazilian buyer who does not specify it contractually will not have it.

The third returns the argument to Axminster. The refusal turned on deflagration mitigation for cabinets spaced 10 cm apart — a layout parameter, and therefore one of installation scale. The developer had compliance and had equipment from a recognized manufacturer. What it did not have was evidence at the level at which the question was asked. There is no indication of bad faith or of product deficiency; there is an incompatibility between the class of evidence available and the claim that needed to be sustained.

Note, finally, that the 2026 edition of NFPA 855 began treating active thermal-runaway propagation prevention systems in a section of their own, defined as automatic systems that detect precursors — gas venting or anomalous temperature — and trigger a directed means of suppression or cooling, with passive measures complementary and insufficient, on their own, to constitute them [16]. This is the perception layer described in the previous paper [3] appearing as a code requirement, which widens the set of claims for which the buyer will have to require evidence of class E3 or above.

VI. Correspondence Rule Between Claim and Evidence

The typology in Section IV only operates when paired with a rule that determines, for each claim, the minimum class required. Table III applies the rule to the specification items proposed in the previous paper [3], with one correction: the first claim decomposes into two, because design and effectiveness under failure require different classes. The class column indicates the floor; sufficiency remains conditioned on criteria (ii) and (iii) of Section IV-B.

Table III — Minimum class required by claim.

Vendor claim	Minimum class	Note
Segregation between the safety function and the learned component was designed	E3	Architectural description identifying the segregated hardware and mapping the principles adopted to the risks and methods described in ISO/IEC TR 5469:2024
That segregation remains effective under failure of the learning layer	E4	A design description does not demonstrate independence under failure; requires testing with fault injection into the inference layer
The fast loop meets the timing envelope	E4	Worst-case execution time under design thermal conditions
Precursor detection is effective	E3	False-negative rate in the operating region, detection latency, and behavior under sensor failure
Thermal propagation is contained	E3; installation scale when the claim involves separation distance or suppression	The test level delimits the sustainable claim
The safety function cannot be disabled	E3	Declaration that no operating, service, or commissioning mode removes it, and the corresponding maintenance procedure
Arbitration among controllers is deterministic and traceable	E4	Emergent property of integration; see Section VII

Vendor claim	Minimum class	Note
Decisions are traceable	E4	Demonstration in test of logging of active constraints and blocked actions, with defined retention
The system operates autonomously without connectivity	E4	Acceptance test with declared duration, state, and reconnection behavior
Interfaces are open and the data belongs to the owner	E3	Interface specification by integration point; a contractual declaration on its own is E0

Two readings follow from the table.

The first is that **five of the ten claims admit support at E3** — designed segregation, precursor detection, propagation containment, non-disablement, and interfaces — that is, with documents that already exist and whose retrieval imposes no new experimental campaign. The obstacle, in those cases, is one of procurement practice and not of testing cost.

The second is that **the five claims requiring E4 are exactly those concerning behavior under failure or under integration** — effectiveness of segregation, timing envelope, arbitration, traceability, and autonomy. They are also the ones that most frequently receive an answer of a lower class, because they require testing with the real component and cannot be satisfied by a product document.

VII. The Chain of Authority Among Controllers

The previous paper proposed a four-layer reference architecture, with exclusive actuation authority concentrated in the shielding layer [3]. That formulation states a correct design rule and is, at the same time, incomplete as a description of a real asset: in a utility-scale system, **the authority to command, limit, block, or interrupt operation is distributed among subsystems from different vendors**, and the safety argument has to cover the boundaries between them. The broader formulation is adopted here, rather than the "write authority" used in [3], because protection relays, interlocks, and fire detection and suppression systems exercise authority through physical outputs, without necessarily writing into the same control domain.

The effective chain approximates: cell sensing → module BMS → rack BMS → master BMS → PCS controller → EMS and power plant controller (PPC) → SCADA → electrical protection → HVAC, fire, and gas detection systems → telecommunications → operator. Each boundary is an integration point, and it is at the boundaries that the failure modes no component test reveals are concentrated.

The elementary case is the command conflict. The EMS requests power; the PCS controller accepts; the BMS detects that the requested C-rate exceeds the thermal limit admissible for the current state of health. The questions the buyer needs to ask are:

1) Which command prevails, and under what rule?
2) How long does the arbitration take to resolve, and what is the worst case of that time?
3) Which variable blocked the command, and was that recorded in a recoverable form?
4) Is the blocked command automatically re-presented, and at what frequency?
5) Can the plant controller override the block? Under what condition, and with what record?
6) Is there a defined degraded mode, and what is the behavior in each degradation?
7) What happens on loss and on restoration of communication between two links of the chain?
8) How is the execution of a *stale command* prevented after reconnection?

None of these questions is answered by product certification, because none of them is a property of a product. **They are emergent properties of integration**, and the minimum class that sustains them is E4 — closed-loop testing with the real controllers, including injection of communication and sensor faults.

This is also why the answer "the system operates locally in the event of communication loss" is insufficient. The follow-up questions are: for how long; with which setpoint; from which state; with which functions degraded; and what happens on reconnection. The absence of those five pieces of information converts an operational claim into a commercial statement.

The implication for specification is direct: **the authority matrix and the arbitration behavior must be contractual documents, produced before the supply arrangement is defined.** After integration, changing them involves the BMS, PCS, EMS, protection, and supervisory interfaces simultaneously, converting an architectural decision into a retrofit.

VIII. A Taxonomy of Non-Response

An answer that does not sustain the claim is rarely an explicit refusal. It is a substitution. Five modes can be distinguished, each revealing something about the architecture or about the engineering process behind the proposal. In the terminology of eliminative argumentation [13], each mode corresponds to a *defeater* that remains uneliminated.

Metric substitution. The answer offers a correlated but insufficient quantity: mean latency in place of worst-case

execution time; aggregate detection accuracy in place of the false-negative rate in the operating region; contractual availability in place of behavior under failure.

The latency case deserves demonstration, because it is the most frequent and the most readily accepted. Consider a loop with a deterministic requirement of 10 ms whose controller exhibits a mean latency of 4 ms, a 99th percentile of 8 ms, and an observed maximum of 47 ms. By the mean, the system operates with 60% margin. By the requirement, the system violates the deadline. Both statements are simultaneously true, and only the second describes the behavior of the control. **Mean latency is insufficient, by construction, to demonstrate compliance with a worst-case timing requirement** — and, when presented as though it demonstrated one, it is a statement occupying the place of a measurement. As argued in the previous paper, an estimator whose inference occasionally takes several times the typical time is not a slow estimator: it is an estimator that breaks the loop intermittently and in a way that is hard to diagnose [3].

Artifact substitution. The answer offers a certificate when a report was requested, a declaration of conformity when data was requested, a commercial presentation when a procedure was requested. What the mode reveals: the report may exist and not have been read internally, or it may contain scope limitations the vendor would rather not expose.

Level substitution. The answer offers evidence of a lower level to a question of a higher level — a module test answering a question about unit-to-unit propagation or about separation distance. It is the hardest mode to detect, because the answer is technically true and formally pertinent. It is the mode that produced Axminster.

Configuration substitution. The answer offers evidence about a variant within the same product family — another chemistry, another module configuration, another enclosure, another suppression system, another firmware version. Any change of cell chemistry, module configuration, enclosure material, or suppression system after the date of the original test demands a new test, regardless of the edition in force [17]. This is a violation of condition (ii) of the sufficiency criterion.

Regime substitution. The answer offers evidence obtained under an operating regime different from the contracted one — temperate-climate data for tropical operation, seasonal dispatch for continuous cycling, arbitrage-neutral operation for multi-service operation, training distribution for operation outside it. This is a violation of condition (iii), and the mode most relevant to the Brazilian fleet entering operation from 2028 onward, for the reasons developed in [3]: experience accumulated under one regime does not automatically transfer to another, and the Victorian Big Battery failure mode occurred precisely at the transition between regimes [18].

The practical rule that follows from the taxonomy is simple to state and uncomfortable to apply: **non-response is information, and it must be recorded as such in the assessment, not treated as a documentary pending item to be settled later.** A pending item settled after signature no longer has the contractual leverage that would make it settleable.

IX. Scope of Validity of Evidence

Conditions (ii) and (iii) of the sufficiency criterion unfold into three dimensions that delimit what a document sustains.

Edition of the method. The procedure under which the measurement occurred defines what was measured. In the UL 9540A case, reports predating the 6th edition were produced under a method that admitted the early termination discussed in Section V, and the new edition takes effect on January 1, 2027 [14]. There is even a proposal that the UL 9540A edition appear on the system's marking plate and in its instructions, which indicates that the normative ecosystem itself treats the edition as identifying information, not as a bibliographic detail.

Configuration identity. The tested object must be the delivered object. Changes of chemistry, module configuration, enclosure, suppression, or firmware break the identity and invalidate the transfer [17]. In procurement practice, this means tying the evidence to the bill of materials and to the design revision of the proposal, and not to the commercial name of the product.

Presumed operating regime. Every measurement embeds hypotheses about the condition of use. The Victorian Big Battery finding regarding the gap between the burn-test envelope — which covers winds on the order of 12 mph — and the gusts of up to 35 mph recorded on the day of the event is the canonical illustration [3], [18]. For Brazil, the critical dimension is thermal: the margin between junction and ambient temperature, which governs the thermal throttling of embedded inference, is compressed by the local climate precisely during the hours of greatest demand on the asset.

From this follows a specification requirement that does not appear in the previous paper's list and that this paper adds: **evidence must come accompanied by a declaration of its scope of validity** — edition of the method, identification of the tested configuration, and presumed operating regime. Without those three pieces of information, conditions (ii) and (iii) are not verifiable, and the document is not assessable, whatever its intrinsic quality.

X. Application: Reading a Constructed Dossier

The dossier that follows is constructed by this author for illustrative purposes. It corresponds to no vendor, product, or real process, and it should not be read as an assessment of any

market participant. The answers were written to represent plausible patterns, not to caricature.

Context. A buyer specifies a 60 MW / 120 MWh system, LFP chemistry, connected to the transmission grid with a grid-forming requirement, under a multi-service regime with market exposure.

Designed segregation. *"The system employs advanced artificial-intelligence algorithms integrated into the BMS, with a redundant architecture and ISO 9001 certification."* — Class E1a. The question was about the authority boundary; the answer describes capability and presents a management-system certification, which says nothing about control architecture. Artifact substitution combined with metric substitution. The absence of any mapping to the methods of ISO/IEC TR 5469:2024 suggests the boundary was inherited from the vendor's platform, not designed for this asset.

Timing envelope. *"Typical controller response latency below 5 ms."* — Class E0. Metric substitution in its purest form, and exactly the case demonstrated in Section VIII. "Typical" is a mean; the claim requires the worst case under design thermal conditions.

Detection operating point. *"Detection accuracy above 98% on a validation set."* — Class E2. Measurement occurred, on the wrong quantity. There is no false-negative rate in the operating region, no detection latency, and no behavior under sensor failure. Given the asymmetric cost between false negative and false positive, and given the Moss Landing precedent, in which the false positive produced the damage itself [3], aggregate accuracy is irrelevant to the decision.

Containment. *"System tested to UL 9540A; test report available under a confidentiality agreement."* — Class E1b as presented, potentially E3 once the report is read. It is the most promising answer in the dossier and the one the buyer should pursue. The questions to ask are three: at what level did the testing stop; under which edition was it conducted; and does the tested configuration correspond to the one offered. If it stopped at module level, the dossier sustains nothing about the proposed separation distances.

Non-disablement. *"Safety functions follow industry best practices and can only be altered by authorized personnel."* — Class E0, and the most worrying answer of the set. "Only by authorized personnel" does not deny the existence of a disabling mode: it asserts access control over it. The question was whether the mode exists. This is exactly the configuration that produced the Victorian Big Battery event, in which protection was removed by switching to service mode and the fault went undetected for thirteen hours [3], [18].

Arbitration among controllers. *"The BMS has priority over the EMS under limit conditions."* — Class E0. The answer states a precedence rule without reporting resolution time, logging of the blocking variable, re-presentation behavior of the command, the possibility of override by the PPC, or the definition of a degraded mode. Of the eight questions in Section VII, it partially answers one.

Traceability. *"The system logs events and alarms with timestamps, exportable via API."* — Class E2. Event logging is not logging of active constraints and blocked actions. The distinction matters under the penalty regime of Resolution 1,162/2026, which brings storage operators under the Brazilian regulator's inspection and sanction regime [19]: what has to be demonstrable is why the system did not execute a given action, and that requires the record of the constraint that barred it, not of the alarm that accompanied it.

Autonomy without connectivity. *"Local operation guaranteed in the event of communication loss."* — Class E0. No declared duration, no setpoint, no starting state, no identified degraded functions, no reconnection behavior, and no treatment of stale commands.

Interoperability. *"Standardized Modbus TCP and IEC 61850 interfaces; operational data accessible through the vendor portal."* — Class E3 as to protocol, E0 as to data ownership. Access through a vendor portal is not data ownership, and the answer does not address the controller integration point, which is where the dependency actually forms.

Consolidated reading. The dossier answers nine of the ten claims in Table III: the effectiveness of segregation under failure receives no specific answer. Of the nine answers, one is an organizational certification unrelated to the product offered, four are statements without evidence, two present measurement of an inadequate quantity, one is partial, and one is potentially sustainable by means of a document that already exists.

No answer reaches E4. Of the five claims that require E4 under Table III, four are explicitly addressed in the dossier and remain unsupported; the fifth — effectiveness of segregation under failure — receives no answer at all. The dossier does not permit forming a judgment about behavior under failure.

The methodological point is that this conclusion was reached **without assessing the product**. It was not asserted that the system is unsafe; it was asserted that the dossier does not demonstrate that it is safe. These are distinct propositions, and only the second is within reach of someone reading documents.

XI. Institutional Use: Three Evaluators, Three Loss Functions

Licensing authority, insurer, and lender read the same dossier with different loss functions, and the misalignment among them is an under-recognized project risk.

The **licensing authority** decides under uncertainty with an asymmetric and irreversible loss: authorizing a venture that fails is public harm, whereas refusing is a private cost. Its rationality favors requiring a high class and, in doubt, refusing. Axminster and Pound Road exemplify the pattern — in both, the decision did not assert that the system would fail, but that it had not been demonstrated that it would not.

The **insurer** decides under a quantifiable and recurrent loss, with an actuarial reading. The survey by insurer QBE of 49 United Kingdom fire brigades found 212 fires linked to photovoltaic systems in 2025, against 91 in 2022 — a 133% increase — while the number of installations grew 52%, with fire spread beyond the originating installation rising from 33 to 88 cases over the same period [20]. Observed incidence grew faster than the number of installations, a signal that merits consideration in risk modeling — although the aggregate data do not permit inferring a causal per-installation failure rate directly, since they normalize neither fleet age, nor typology, nor exposure, nor installation quality, nor any changes in reporting criteria. What the series does sustain is that the hypothesis of stable per-unit claims experience cannot be assumed without verification, which pushes underwriting toward evidence of a higher class, or toward a premium that substitutes for it.

The **lender** decides on cash flow over a contracted horizon, with a double exposure: the safety event and the inability to stack revenues. In the 30 MW/60 MWh asset on the São Paulo transmission system — the Registro substation installation, Brazil's first utility-scale grid-connected storage asset — the published study indicates that extending it to certain additional services may require changes to protection settings, control algorithms, and auxiliary software and hardware [21], which demonstrates that nameplate capacity, on its own, does not define the functional aptitude of the asset. Control architecture is therefore a variable of financial modeling, and not only of safety.

The misalignment risk is concrete: a dossier sufficient for technical qualification may be insufficient for environmental licensing, and a dossier sufficient for both may not sustain the revenue hypothesis underpinning the financing. Since the three assessments occur at different moments and the contractual leverage over the vendor expires at signature, **the class of evidence to specify is the one required by the most demanding evaluator in the chain, and not by the first barrier to be cleared.**

XII. What This Analysis Does Not Demonstrate

This analysis sustains that the distinction between classes of evidence is operationalizable; that evidentiary sufficiency requires, simultaneously, adequate class, configuration identity, and regime applicability; that the stopping logic of UL 9540A produces, by construction, dossiers silent as to installation scale, and that this gap was recognized and corrected in the codes in 2026 in the United States and remains open in the Brazilian framework; that authority arbitration among controllers is a claim distinct from the others and requires a class of evidence of its own; and that non-response admits classification and is informative.

It is **not** a validated methodology, and this is the principal limitation of the work. The dossier in Section X was constructed by this author, with answers selected to illustrate substitution modes defined by the author himself. That exercise demonstrates that the instruments produce an internally consistent reading; it does not demonstrate that they produce a correct reading, nor that they discriminate adequately between dossiers of differing quality. Validation would require application to a set of real anonymized dossiers, with independent assessment of the results — work that has not been carried out and that constitutes the natural continuation of this line.

The typology in Table I is a conceptual proposal, without quantitative calibration. Classes E0 to E3 are ordered by recoverable information; above E3, comparison depends on the two axes described in Section IV-A and admits no single ordering. No metric of relative strength between E4 and E5 is proposed.

The taxonomy in Section VIII is neither exhaustive nor derived from a systematic sample of market responses. Additional modes probably exist, and the relative distribution among them is unknown.

Nor does this analysis constitute an assessment of any vendor, product, or technology, and it does not replace project risk analysis, hazard mitigation analysis, or a conformity opinion issued by a qualified professional. It does not address the legal or contractual aspects of procurement, which require a different competence.

Finally, the instruments proposed here address the reading of evidence that is presented. They do not resolve the case in which the necessary evidence does not exist — a situation that, for the characterization of worst-case execution time under thermal throttling in a container operating in a tropical climate, remains the most relevant experimental gap for the Brazilian fleet entering operation from 2028 onward [3].

XIII. Conclusion

The previous paper closed by observing that the window for correctly specifying the assets of the Brazilian storage fleet closes before they exist. This one adds that specifying without knowing how to read produces the same exposure, with more paper.

The asymmetry that separates buyer from vendor survives even when the information exists and is disclosed: it is one of competence to assess scope. It is not resolved by demanding more certificates, because a certificate is a conclusion detached from its premises. It is resolved with four

instruments, all of immediate application: a typology of evidence classes; a sufficiency criterion that requires, simultaneously, adequate class, compatible configuration, and applicable regime; an explicit matrix of authority and arbitration among controllers; and the habit of recording non-response as information, and not as a pending item.

The proposition that organizes them is that there is no strong evidence in the abstract. There is correspondence, or absence of correspondence, between the claim to be sustained and the evidence available — a correspondence that depends simultaneously on configuration fidelity and on the realism of the condition under which the observation was made.

The intervention with the best ratio of cost to evidentiary gain is modest and almost never practiced: ask for the report instead of the certificate, and require that it come with level, edition, and configuration declared. It imposes no new experimental campaign; it recovers evidence already produced, even if it may call for a confidentiality agreement.

Axminster demonstrated that compliance achieved does not license. Pound Road demonstrated something harder: not even the absence of objection from the technically competent authority licenses, if the demonstration is not made. Between the technical qualification in November, the auctions in December, and the start of supply in August 2028, every dossier read in Brazil will be subject to the same criterion — with the difference that here the regulatory gap NFPA 855 closed in 2026 remains open, and what is not required in the contract will not be obtained afterward.

REFERENCES

[1] Energy Research Company (EPE), *Registration Results — LRCAP 2026 Storage*. Rio de Janeiro, Brazil: EPE, Aug. 2026.

[2] Brazil, Ministry of Mines and Energy, *Normative Ordinance No. 136, of June 1, 2026*. Establishes the guidelines and procedures for the Capacity Reserve Auction through new battery energy storage systems — LRCAP 2026. Brasília, Brazil: Official Gazette of the Union, 2026.

[3] E. M. Fagundes, "Compliance is not safety: a demonstrable safety argument for energy storage systems and the control architecture that sustains it," Zenodo, Sep. 2026, doi: 10.5281/zenodo.22310318.

[4] United Kingdom Ministry of Defence, *Defence Standard 00-56: Safety Management Requirements for Defence Systems*. Glasgow, UK: MoD.

[5] T. P. Kelly and J. A. McDermid, "Safety case construction and reuse using patterns," in *Proc. 16th Int. Conf. Computer Safety, Reliability and Security (SAFECOMP)*, York, UK, 1997.

[6] Assurance Case Working Group, *Goal Structuring Notation Community Standard*, current version. Safety-Critical Systems Club.

[7] International Organization for Standardization and International Electrotechnical Commission, *ISO/IEC 15026: Systems and Software Engineering — Systems and Software Assurance*. Geneva: ISO/IEC.

[8] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, *Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)*. York, UK: Assuring Autonomy International Programme, University of York, 2021.

[9] UL Standards & Engagement, *UL 4600: Standard for Safety for the Evaluation of Autonomous Products*. Northbrook, IL: ULSE.

[10] International Organization for Standardization and International Electrotechnical Commission, *ISO/IEC TR 5469:2024: Artificial Intelligence — Functional Safety and AI Systems*. Geneva: ISO/IEC, 2024.

[11] Sandia National Laboratories, *Analyzing System Safety in Lithium-Ion Grid Energy Storage*. Albuquerque, NM: Sandia National Laboratories, 2016. [Online]. Available: https://www.osti.gov/servlets/purl/1257985

[12] R. Bloomfield and J. Rushby, "Assurance of AI systems and the confidence in assurance cases," technical report, 2020.

[13] J. B. Goodenough, C. B. Weinstock, and A. Z. Klein, "Eliminative induction: a basis for arguing system confidence," in *Proc. 35th Int. Conf. Software Engineering (ICSE)*, San Francisco, CA, 2013.

[14] UL Solutions, "UL 9540A test method for battery energy storage systems," and "Understanding UL 9540A, NFPA 855 and large-scale fire testing for battery energy storage systems," Northbrook, IL, 2026.

[15] National Fire Protection Association, *NFPA 855: Standard for the Installation of Stationary Energy Storage Systems*, 2026 ed., Annex G.11. Quincy, MA: NFPA, 2026.

[16] Telgian Engineering & Consulting, "NFPA 855 changes in the 2026 edition," 2025.

[17] UL Solutions, "Installation codes and requirements for energy storage systems (ESS) — FAQs," 2026.

[18] Energy Safe Victoria, *Statement of Technical Findings: Fire at the Victorian Big Battery*. Melbourne, Australia: ESV, Sep. 28, 2021.

[19] Brazilian National Electric Energy Agency (ANEEL), *Normative Resolution No. 1,162, of June 2, 2026*. Amends Normative Resolution No. 1,000 of December 7, 2021, to regulate billing and the regulatory classification of electric energy storage systems. Brasília, Brazil: ANEEL, 2026.

[20] QBE Insurance, "Solar panel fires increasing at more than double the rate of installations," QBE European Operations, London, UK, Aug. 2026.

[21] P. F. Torres, A. R. A. Manito, G. Figueiredo, M. P. Almeida, J. C. S. Almeida Neto, R. L. Cavalcante, C. C. V. F. A. Silva, and R. Zilles, "Energy storage as a transmission asset — assessing the multiple uses of a utility-scale battery energy storage system in Brazil," *Energies*, vol. 18, no. 4, art. 902, 2025, doi: 10.3390/en18040902.