Temporal auditing of AI outputs

Rubén Abella · SPE Audit
Research edition 1.2 · 22 November 2026

Abstract

Four events from space missions underpin the original comparison. The 96 development responses combine four events, eight formulations and three reporting cutoffs. Another 16 responses vary the wording while retaining the same events. We ask whether a statement presents an outcome after the cutoff as already accomplished and what evidence an application can preserve for someone else to review that judgment.

The rule supplied with outcome dates detected 24 of 32 positives in development. With new wording it detected 0 of 8 and referred 14 of 16 cases. Its two automated decisions were correct. A reviewer receiving those 16 responses would still have to examine the other fourteen.

In this edition a local model read the same 112 responses with and without a field giving the outcome date. Adding it changed six development decisions and none in the challenge. On the challenge both passes detected **4/8** positives, raised one false alert and referred **11/16** cases.

The author develops the technology under evaluation. The project used AI assistance in preparing texts, rules, analyses and documentation. Classification references are editorial judgments. The human review pilot remains open.

Version DOI · 10.5281/zenodo.24447732

What the answer asserts and what the record preserves

Before DART hit Dimorphos, the plan to attempt the impact was public. At a 1 January 2022 cutoff, “an impact is planned” describes an intention. “DART hit Dimorphos” presents the outcome as accomplished. NASA reported the impact of 26 September 2022. The difference depends on the verb, the referent and the cutoff, even if both answers give the same year.

Term	Meaning in this study
Reference positive	The answer asserts that the event after the cutoff has already happened.
Reference control	The answer expresses a different proposition, such as a plan, a denial or an event completed by the cutoff.
Alert / Other / Review	Reader outputs. Review defers the decision. An indeterminate reference records uncertainty in the reference label itself.

The capture boundary

The adapter records the request bytes delivered to the local sender, the returned text and their digests. A reviewer can recover the question, cutoff and eligible source cards and check whether they match a retained record.

Suppose two services return the same sentence. One repeats a forecast that happens to be right. The other retrieves a later news report outside the application. If the client observes the same bytes in both cases, it also obtains the same digests. Attributing the answer to a source would require an observation that distinguishes the two paths.

The original **124 captures** comprise 96 development responses, 16 new-wording responses and 12 checks of local source selection. In these last checks a function returned the text of the latest eligible card. Removing that card tested the function’s behaviour. Only the first 112 responses enter the classification results. The AI-assisted editor who wrote the 16 challenge texts knew the research question and the general limit of a year rule. The editor fixed the texts before inspecting the date-aware rules or development examples.

Reading with and without the outcome date

The same local reader classified each answer twice. Both passes receive the inclusive cutoff, question, answer, and the date and text of each eligible source card in recorded order. URLs and descriptive card identifiers are removed. One pass adds a single field giving the registered outcome date, which can fall after the cutoff. The other omits that field.

Method	Information available for the decision
Year rule 1.0	Explicit years in the answer and the cutoff year. The original experiment sets the future-context signal to false.
Date-aware rule 1.0	Answer text, cutoff, registered event date and outcome phrases chosen for each mission. The rule can abstain.
Local model 1.2	Identical model and inputs in both passes apart from the date field. The template permits an alert, other or review.

Model and execution. SmolLM2-135M-Instruct, Q4_K_M weights, llama.cpp b8058, CPU, temperature 0 and seed 20261122. Each request starts without chat history. The first pass alternates by row. These are 224 inferences about previously written answers. Reference labels are joined by identifier at scoring time. The original answers were written separately from this inference.

The corpus and earlier results were already public. The prompt, inputs, code and their digests were fixed before this run. DigiCert issued an RFC 3161 timestamp for the plan digest on **22 November 2026 at 19:23:03 UTC**. The timestamp dates the commitment to this plan. Labels still come from the project. Timing may also be inferable from eligible cards, the answer or information encountered in model training.

`SHA-256 9402541a361800c9b2f7dddf1ccdb36c8cd6e383c7b46c6a5838f34c0252073e`

Frozen plan · Exact instructions and verifiable timestamp

Decisions and review workload

The tables retain their denominators. **Detected** counts correct alerts over all positives. **False alerts** are counted over all controls. **Precision** is the proportion of alerts that are correct. **Coverage** counts automated decisions over all rows. **Review** counts abstentions over those same rows.

96 development responses · 32 positives, 64 controls

Method	Detected	False alerts	Precision	Coverage	Review
Year rule	4/32	12/64	4/16	96/96	0/96
Date-aware rule	24/32	0/64	24/24	84/96	12/96
Model without date	8/32	16/64	8/24	55/96	41/96
Model with date	8/32	17/64	8/25	54/96	42/96

16 new-wording responses · 8 positives, 8 controls

Method	Detected	False alerts	Precision	Coverage	Review
Year rule	1/8	2/8	1/3	16/16	0/16
Date-aware rule	0/8	0/8	—	2/16	14/16
Model without date	4/8	1/8	4/5	5/16	11/16
Model with date	4/8	1/8	4/5	5/16	11/16

Correct among automated decisions, in table order: 56/96, 84/84, 29/55 and 27/54 in development. In the challenge: 7/16, 2/2, 4/5 and 4/5. All 224 model receipts yielded a valid label. Alert precision for the date-aware rule on the challenge is undefined because it raised no alerts.

The 12.5% sensitivity of the year rule and 75% sensitivity of the date-aware rule in development reflect differences in both design and access to information. In the challenge, 2/2 correct automated decisions by the date-aware rule cover only 2/16 responses. All eight positives remained open. These fractions cannot estimate a production error rate.

Row-level decisions · Counts by event · Transitions by reference label

Why the individual missions matter

On the 96 development responses, both model passes detected eight of the 32 positives. All eight belonged to Perseverance. The model raised an alert on every one of the 24 answers for that mission, including all 16 controls. Across the other three missions, it missed or referred every development positive.

Mission	Positives detected without / with date	False alerts without / with date	Referred without / with date
Perseverance	8/8 · 8/8	16/16 · 16/16	0/24 · 0/24
DART	0/8 · 0/8	0/16 · 1/16	17/24 · 18/24
OSIRIS-REx	0/8 · 0/8	0/16 · 0/16	0/24 · 0/24
Rosetta	0/8 · 0/8	0/16 · 0/16	24/24 · 24/24

The six decisions that changed were DART controls. Three moved from other to review, two from review to other and one from other to alert. The same number of detected positives therefore coexisted with one additional false alert and a different set of cases awaiting review.

On the 16 new-wording cases both passes made the same 16 decisions. Of the four detected positives, two concerned Perseverance and two concerned DART. Rosetta and OSIRIS-REx together contributed eight responses. The model referred all eight. Each mission accounts for just two reference positives in this set.

Counts by mission and method · Paired decisions

Eight decisions open to inspection

These are the eight cases selected for the original paper. The final two columns add the measured 1.2 model decisions. **P** = reference positive or detector alert, **O** = other and **R** = review. “Without” and “with” refer to the registered outcome-date field in the local model input. Readers classified the English texts.

Case	What the answer does	Ref.	Year	Dated	Without	With
perseverance-04	Forecasts landing	O	P	O	P	P
perseverance-05	Denies confirmation	O	P	O	P	P
osiris-06	Quotes a question	O	P	O	O	O
rosetta-08	Affirms wake-up	P	O	P	R	R
dart-09	Affirms impact	P	O	P	R	R
h03	Poses a hypothetical	O	O	O	R	R
h10	Reports an earlier fact	O	O	R	R	R
h16	Affirms “did so”	P	O	R	P	P

When a pronoun changes the reading

“Perseverance was built to land on Mars. It did so in February 2021.” The cutoff for **h16** is 10 February. The first sentence explains what the rover was built to do. The second picks up the verb and asserts that it landed. The landing occurred on the 18th. The date-aware rule referred this case. By contrast, “There is no confirmation that Perseverance landed on Mars in 2021” is a control at the 2020 cutoff. The year rule raised a false alert on the future year.

The eight examples illustrate particular disagreements. The public decision file contains all 112 responses for readers who wish to count how often an outcome occurred within this study.

Development texts · Challenge texts

Evidence beyond the four missions

Six new events, 48 responses in English and Spanish

Supplement 1.1 assessed 18 positives, 24 controls and six indeterminate references about transport, the performing arts and experimental physics. The proposition rule detected **1/18**, raised one false alert on 24 controls and referred **24/48** cases. In v038 the speaker repeats the premise that an inauguration took place, then rejects it. The rule picked up the assertion in the first clause.

An exploratory AI-assisted reading matched the 18 positives, 24 controls and six ambiguities. It was commissioned after the aggregate rule results were known. A later follow-up used a local reader with fixed instructions and weights. It detected **11/18** positives, raised alerts on **13/24** controls and referred **17/48** cases. Its decisions repeated on a second pass. The protocols differ, as does what each reading establishes.

Seventeen passages written by others

Seventeen English passages published by NIST and the USGS concern six events. The agencies published the wording. The project selected the passages, assigned historical cutoffs and wrote the questions and reference labels. The fixed local reader in the follow-up detected **0/5** positives and referred **16/17** passages. Its one automated decision closed a positive as other.

This purposive collection diversifies the source of wording. Reference judgments still originate within the project. The separate four-language translation replay reuses the original answers. None of these datasets measures how often a remote service actually asserts a post-cutoff event.

Supplement 1.1 · Follow-up results and institutional sources

Prior art and the strength of the record

Four works locate the contribution. The review deliberately sought specific technical operations and identifiable versions. The separate catalogue retains cross-references, search leads and records whose historical origin remains unresolved.

Precedent	Relevance to this question
Dantart · N01	Historical retrieval and a signed per-query record covering documents, response and time policy.
EQTY Lab · N02	Cryptographic evidence linking retrieved data to an inference input. The attested boundary determines the scope of the claim.
Liu et al., ExAnte · N03	Before-and-after reference sets for assessing assertions whose verifiability depends on a date.
Paleka et al. · N06	Earlier pages updated with the outcome and search biases that can survive a nominal date restriction.

SPE Audit places the retained request, the applied rule and the observed disagreement in a single record that another person can review. The **116-reference** catalogue is issued separately. Administrative and commercial items retain their identifiers and documentary purpose. Version 1.1 revisited 58 entries whose dating was initially inadequate. An identifiable object could be dated in 53. Five remained unresolved. A snapshot or update dates a particular version.

In a dispute. The record documents the disputed answer, the cutoff and retained cards. The reviewer must distinguish occurrence, publication, entry into force and capture. Articles 12 and 14 of the EU AI Act frame questions about record-keeping and human oversight for high-risk systems. Regulatory conformity requires a deployment-specific assessment.

Prior-art catalogue · Dating register

Independent review and replication

34 candidate events. The register for the next study records 28 realised events, five documented cancellations and one postponement. Source pages are in English (19), Spanish (10), French (4) and German (1). Selection was purposive. Several events may share a common cause. The sources document events and dates. The questions, evaluation references and predictions for this cohort remain to be produced.

A Canary Islands record illustrates a choice the authors of those questions will face. The César Manrique Foundation formally inaugurated its headquarters in Tahíche on 27 March 1992 and opened to the public on 1 April. At a 30 March cutoff, “was inaugurated” and “opened to the public” require different judgments. Merging the two milestones would put the error in the reference before any model is evaluated.

Open human pilot. Participants can use the 16-case kit to obtain two separate decisions per case and ask a third reviewer to resolve differences. Prior exposure to cases is recorded. At release the project has measured zero human decisions and zero minutes of human review. The published data cannot estimate inter-rater agreement or the monetary cost of a referral.

Reproduce the tables. From v12, run `python -B score.py` with Python 3.11 or later. The program checks the plan, reads the archived receipts and recomputes decisions and counts. The repository also supplies code and instructions for a fresh local inference using the published weights. An external replication can report the version, plan digest and any differences by case identifier.

Replication materials · 34-event register · Open pilot

Citation. Rubén Abella (2026). *Temporal auditing of AI outputs*. SPE Audit, research edition 1.2. DOI 10.5281/zenodo.24447732. Original text and figures under CC BY 4.0. Code is distributed under the repository’s BSL 1.1 licence. Trademarks and third-party materials retain their own terms.