SERIES

THE RECORD · FOUR PAPERS · OPEN LICENCE

The Measured *Endpoint*.

Working notes on scoring live model endpoints: who judges the output, and what makes a measurement evidence

PAPERS

ISSUED

four papers · published 14 September 2026 · CC BY 4.0

digital1.foundation/articles/the-measured-endpoint/

Stichting Digital One Foundation i.o. · Amsterdam

These notes describe work contributed to an entity in formation.

TOC

THE MEASURED ENDPOINT · FOUR PAPERS · OPEN LICENCE

Contents

001

PAPER 001 · METHOD

A judge is a *party to the score*.

Free-form model output is scored by another model, and models used as judges have shown measurable preferences, including for text like their own: where a judge's vendor also makes models it scores, the judge is a party to the score. A panel drawn from several providers dilutes that preference without removing it, and a rule that rejects the outlier among few judges — no panel size is published — protects at most against one member failing alone, not against anything two members share. These notes publish no rubric, prompt set or panel composition, so no score can be re-run from them.

PAPER · 001 PUBLISHED · 14 SEP 2026 READING · 7 MIN

Canonical · digital1.foundation/articles/the-measured-endpoint/001-a-judge-is-a-party-to-the-score.html

§1

Free-form output needs a judge, and the judge has a stake

A model asked a question with one correct answer can be scored by comparison. Asked to explain, plan or write, it produces text that no answer key matches, and scoring that text hour after hour, across model series from several providers, is not work people can do at the rate an endpoint changes. So a model scores it. That moves the question rather than answering it: the measurement is now partly a property of the model doing the measuring.

The literature is specific about how. Zheng and colleagues, examining models used as judges, name position bias, self-enhancement bias and verbosity bias — an LLM judge that "favors longer, verbose responses, even if they are not as clear, high-quality, or accurate as shorter alternatives" — alongside limited reasoning ability.[1] On self-enhancement their own result is inconclusive: "our study cannot determine whether the models exhibit a self-enhancement bias."[1] The firmer evidence came later. Panickssery, Bowman and Feng describe self-

preference, "where an LLM evaluator scores its own outputs higher than others' while human annotators consider them of equal quality", and find its strength tracks a model's ability to recognise its own text.[2] Their setting is summarisation with three models, and they call their causal case evidence rather than proof.[2] Neither measures a current judge, so neither sizes the effect in one.

> THE PREMISE, PRECISELY
>
> Where the vendor of a judge also makes models among those the judge scores, the judge is not outside the measurement. It is ***a party to the score***, and a method that uses it has to say how that is contained, and where the containment stops.

What follows is the part of that method which answers the stake — a panel and a rule for rejecting its outliers — each with the limit that belongs to it. The design is an operator's, named at the end. Nothing on this page meets the condition this record has already named for a score's soundness to be answerable, and the last section says so before it is asked.

§2

A panel from several providers dilutes a preference

The design describes a panel: judge models drawn from different providers, each given the same transcript and the same rubric, each returning integer quality and reasoning scores from 0 to 100, which an aggregator combines. As designed, no single vendor's judge decides an aggregate alone. There is published support for the shape. Verga and colleagues had "each evaluator model independently" score an output, pooled the scores, and found that in their settings a panel of smaller models from disjoint families outperformed a single large judge in agreement with human judgements and "exhibits less intra-model bias due to its composition of disjoint model families".[3]

Their bounds travel with the result: three evaluator settings, a three-member panel, pooling by max voting or by average, and no outlier rejection, with the choice of which models belong on a panel left "to future work", and the authors work for the maker of one of the three panel members.[3] The same paper reports what the premise predicts: "the highest positive delta for each individual model being scored occurs when it is judged by itself."[3]

So a panel does not remove a judge's preference. It dilutes it, and unevenly. If members come from providers whose models are also scored, a provider with a judge on the panel has its models scored by a panel

containing at least one member of its own family, and a provider with none does not. Averaging shrinks that family's weight; it does not make the aggregate neutral, because neutrality is agreement with a reference, and nothing these notes could read names one. The design does not say whether a member scores its own vendor's models, or how such a verdict is marked.

> INDEPENDENCE IS ASSUMED, NOT SHOWN
>
> Judges from different vendors are not thereby judges with independent errors. Models trained on overlapping public text can share a blind spot, and a panel that agrees has shown agreement, not correctness. These notes hold no measurement of how correlated the panel's errors are.

As designed, provider diversity also buys fault tolerance: when one provider is down, the panel is meant to lose a member rather than fall silent, and nothing in the public delayed feed (read in Paper 002) shows which happened in any given hour. That is where the instrument changes. A panel short of a member is a different estimator, so a score series running through an outage mixes two, and nothing a reader receives marks the window.

§3

Rejecting the outlier among few judges is a policy

The design rejects outliers per verdict: a member whose scores disagree with the rest is excluded before aggregation, and the exclusion is recorded. The statistic, the threshold, the minimum number of members that must survive, and whether quality and reasoning are tested separately or together are not described.

The size of the panel decides what such a rule can mean. Take three, the smallest panel in which one member can be outvoted — an illustration, since no panel size is published. Among three, "outlier" is defined only by the other two, so rejecting the most distant member hands the aggregate to the two closest. That protects against one member failing alone: a truncated reply, a refusal read as a zero, a judge misreading the rubric. It protects against nothing two members share, and where two are wrong in the same direction it can discard the one that was right.

Statistics said as much long before models judged anything. The NIST/SEMATECH handbook warns that Z-scores "can be misleading (particularly for small sample sizes)", that "we typically do not want to simply delete the outlying observation", and that deletion is warranted only "if it can be determined that an outlying point is in fact erroneous".[4]

The authors it cites recommend that modified Z-scores with an absolute value above 3.5 "be labeled as potential outliers" — labelled, not removed.[4] The section assumes roughly normal univariate data, which integer scores from a handful of judges are not, so it supports a caution and endorses no rule. In that illustration, with two members left — one provider down — there is no outlier to define at all, and the safeguard is off when the panel is weakest.

What makes the rule defensible is not the rejection but the flag. An excluded member stays in the record with `rejected` set, so whoever holds the record can ask afterwards, for as long as rows are retained (a period the design does not state), how often each member was overruled, and on whose models. What that record holds, and what it cannot show a reader, is the subject of Paper 002.

§4

What a panel cannot show a reader

The Delegation Record's Paper 007 names what would make a score's soundness answerable, "a published method someone other than its author can re-run, or … outcomes measured on a named corpus";[5] and that condition is unmet here.

Neutrality is not measured. No agreement between the panel and a named human-labelled or keyed reference is described, so whether the aggregate is free of vendor preference is a measurement these notes do not have.

The panel is unnamed. Its composition is not published, so that its members come from different providers is the operator's statement and not something a reader can check. There is a stated reason to withhold it — a named judge is a target an endpoint could be tuned to please — and a shape already in this record that keeps the reason while allowing the check: a commitment by hash,[6] made to each composition when it takes effect and revealed after a stated delay. The design describes nothing of the kind.

> ON THE PROVENANCE OF THIS MATERIAL
>
> The architecture described here comes from ***FrontierScore***, a Digital One property, at frontierscore.ai; it is the provenance of this method, not its subject. Its repository is not public. The design description these notes draw on is its builders' own, and no part of it is checkable today, which is why its scale figures are left out. Deliberately absent: prices, plans and tiers, availability, and deployment internals.

References

1 L. Zheng et al., ***Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena***, NeurIPS 2023 Datasets and Benchmarks Track, arXiv:2306.05685v4, abstract and §3.3; judges studied are 2023 models. arxiv.org/abs/2306.05685, read 14 September 2026.

2 A. Panickssery, S. R. Bowman, S. Feng, ***LLM Evaluators Recognize and Favor Their Own Generations***, arXiv:2404.13076v1, abstract and limitations; summarisation tasks only. arxiv.org/abs/2404.13076, read 14 September 2026.

3 P. Verga et al., ***Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models***, arXiv:2404.18796v2, abstract, §2.2, §3.1 and §4.4; the panel had three members and used max voting or average pooling, with no outlier rejection. arxiv.org/abs/2404.18796, read 14 September 2026.

4 NIST/SEMATECH, ***e-Handbook of Statistical Methods***, §1.3.5.17 "Detection of Outliers"; the modified Z-score recommendation is Iglewicz and Hoaglin's, as summarised there. itl.nist.gov/div898/handbook/eda/section3/eda35h.htm, read 14 September 2026.

5 The Delegation Record, Paper 007, A switch attests what it decided, not that it decided well; its provenance note records that frontierscore.ai's public pages linked no methodology, read 12 September 2026.

6 The Project Handshake, Paper 002, Two trees.

002

PAPER 002 · METHOD

A verdict can be kept, *not re-run*.

As designed, a per-judge record, a content-keyed cache and a separate scoring plane let the operator take an aggregate apart, replay identical output and keep the rubric from workers. None of the three gives a reader of the feed anything to check them with: a cached verdict is repeatable, not reproducible, and the public feed carries values and none of what they were made from.

PAPER · 002 PUBLISHED · 14 SEP 2026 READING · 8 MIN

LICENCE · CC BY 4.0

Canonical · digital1.foundation/articles/the-measured-endpoint/002-a-verdict-can-be-kept-not-re-run.html

§1

Keeping each member's verdict makes the panel legible

The design these notes draw on, named at the end, scores free-form output with a panel of judges whose shape and limits are the subject of Paper 001; this paper follows what happens to a verdict after it. The design persists each aggregate with its members. Per member it stores the judge identity, pinned to an exact model string; its quality and reasoning scores; whether it was rejected; its latency and its token counts. Panel composition carries a version, and a verdict names the composition that produced it. A rubric is bound to a test version: changing how a test scores requires a version bump, so, where the rule is followed, a verdict stays attributed to the rubric it was made under. A rule records the changes someone remembers to bump, and the cache key below names a path by which it could be bypassed.

```
one verdict, as the design describes its fields · no rows are published
aggregate     quality, reasoning
composition   version
test          id, version        the rubric is bound to it
member[ ]     model              an exact model string
              quality            integer, 0–100
              reasoning          integer, 0–100
              rejected           true | false
              latencyMs, tokens
```

Keeping the members is a storage decision, and without it nothing below is possible: an aggregate stored without its members cannot be taken apart later. It is also a record for the party holding the store. The record is the operator's own account of its own panel, the position The Delegation Record calls self-attestation, where labelling "does not lift the record out of that position; it makes the position legible".[1] A store its operator can edit or truncate says nothing about rows no longer in it; of a verified chain, The Project Handshake puts it as "the one attack it structurally cannot see is omission".[2] The design describes no commitment of verdict rows to anything the operator cannot rewrite.

Versioning has a boundary of the same kind. A version bump records a change the operator makes to a rubric. It cannot record a change in a judge's own behaviour under an unchanged model string, because the field naming the model is filled in at the far end, by the party serving it, and nothing between that party and the record checks it;[3] a judge that drifts under a pinned name moves scores while no version moves. And a version shows that something changed, not that the rubric measures what its name says. An unpublished rubric is as uncheckable at a late version as at its first.

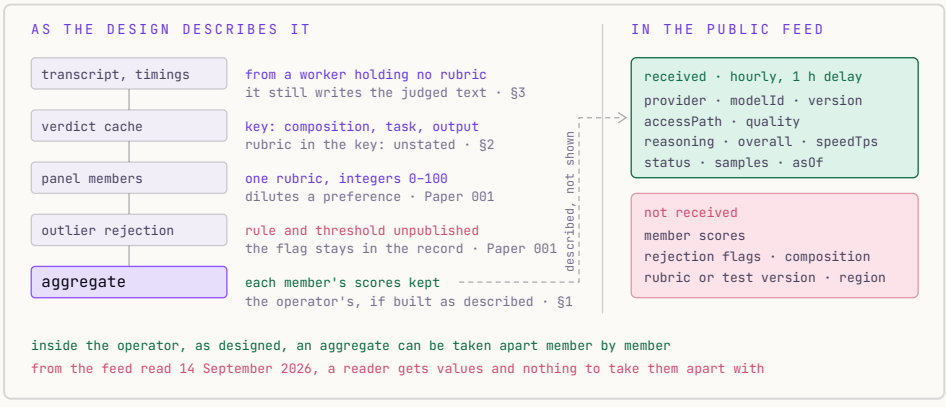

FIGURE 1 The decomposition exists on one side of a boundary. As designed, the operator can take a verdict's aggregate apart member by member. A reader of the feed gets quality and reasoning values over an unstated window of samples, which the operator describes as built from such verdicts. Nothing received shows how, or carries a member record to check it with.

§2

A cached verdict is repeatable, not reproducible

The design keys verdicts on content. A verdict is stored under panel composition, task hash and output hash, and, as described, the cache is consulted before a judge is called. When an endpoint returns output byte-identical to output already judged for the same task by the same composition, the stored verdict comes back and no judging runs. That saves inference on repeated output, and it keeps the record consistent: identical output, identical verdict, for as long as the entry exists.

It does not make the judges deterministic. Atil and colleagues report that "outputs can vary for the same inputs under settings expected to be deterministic", and that among the models they tested "none of the LLMs consistently delivers repeatable accuracy across all tasks, much less identical output strings".[4] Their experiments measure task accuracy on five models, not rubric scoring, so they establish that the variance exists and say nothing of its size in a judge. What the cache holds is one draw from a judge that might have drawn otherwise. Served again, it hides the spread a second judgement would have shown, and a wrong first verdict is replayed to each later identical output. Re-judging a sample of cached keys, and publishing how often a fresh panel disagrees, would size this; no such check is described.

The key is also only as current as its fields. Composition is in it, so a changed panel judges afresh. The rubric version is not among the fields as described, and whether the task hash covers the rubric, the judge prompt template, the decoding settings or the rejection rule is unstated. If it covers none of them, a scoring change that leaves composition alone could be answered with a verdict made under the old scoring. That is an open question about the design, not a defect found in it; nothing available to these notes settles it.

§3

A worker holds no rubric and still writes what is judged

The design runs judging only in a central scoring plane, apart from the fleet that measures. A probe worker submits a raw transcript and its timings; as designed, it holds no rubric and no scoring authority, and cannot edit a stored verdict or the rule that produced one. The fleet, and the checks a result passes before it reaches the panel, are the subject of Paper 004.

Isolation narrows what a dishonest worker can do; it does not take the worker out of the score. The worker supplies the thing that is judged. It can fabricate a transcript, truncate one, or add text addressed to the judges rather than to the task, and a judge reading the transcript reads

that too. The endpoint's provider is outside the isolation as well, since the output is its to shape. The worker also supplies the timings, which no judge reads. For a partner worker's result, what the design puts between a dishonest submission and a score is admission, not isolation. In the design, a first-party result is the baseline others are checked against and is not staged, so a compromised first-party key is covered by neither. Paper 003 treats a measurement as a signed claim and says what a signature does not prove.

Isolation also moves trust rather than removing it: the party that can change a rubric, a panel, a threshold or a cache entry is the operator, in one plane, and nothing published shows a change made there.

§4

What a reader cannot check today, conceded first

These notes describe an architecture. They publish no rubric, no prompt set, no panel composition, no test suite, no aggregation or rejection rule and no calculation for any published field. The Delegation Record's Paper 007 names what would make a score's soundness answerable, "a published method someone other than its author can re-run, or … outcomes measured on a named corpus",[5] and nothing here meets it.

What a reader can check is the public delayed feed, and it shows little. Read 14 September 2026, `/v1/scores` lists 18 model series from five providers — anthropic, deepseek, google, openai and xai — each on access path `api`, delayed one hour at hourly resolution; `/v1/history` holds the same 18 series as 2,938 hourly points over seven days.[6]

OPEN PROBLEM · IDENTIFIED BY READING THE PUBLIC DELAYED FEED AGAINST THE DESIGN, 14 SEPTEMBER 2026

Nothing in the public feed lets a reader recompute a score

The feed a reader receives carries values the operator describes as panel scores, and none of what they were made from. One series, every field as published:

```
/v1/scores · one of 18 series · read 14 September 2026
printed for its fields: identity and score values withheld
{"provider":"<provider>","modelId":"<model id>",
 "version":"<model id>","accessPath":"api",
 "overall":<number>,"quality":<one decimal>,"reasoning":<one decimal>,
 "speedTps":294.3,"status":"healthy","samples":371,
 "asOf":"2026-09-14T13:26:32.133Z"}

absent   member scores · rejection flags · composition
         rubric or test version · region · signature
version names the model's version, not the test's
```

Identity and score values are withheld: printed beside a named model they would read as a measurement of it, and they are the output of an unpublished method for which, for the reason The Delegation Record 007 gives, these notes claim no soundness. The shape is the point.

A reader holding this can see what was published and when. They cannot take `quality` apart into verdicts or members, cannot tell which composition or rubric version produced it, and cannot recompute `overall`. The operator's homepage, read 14 September 2026, lists quality, speed and "an overall preferred score" as separate outputs,[7] and no calculation for `overall` is published in the feed or on the pages read for The Delegation Record 007. Across the 18 series of that snapshot, `overall` correlates at 0.733 with `speedTps` and at −0.308 with `quality` (Pearson, n = 18, computed while assembling these notes). That describes one snapshot, not a formula, and nothing read shows `overall` to be the judged number. The history has the same limit over time: 231 of its 2,938 points carry quality and reasoning of zero, and nothing in the feed separates a panel that scored zero from no verdict at all. What would close it: an export of verdict rows for a named window, a composition and rubric version on each point, and the calculation of each published field.

1 **Withholding prompts cuts both ways.** A published prompt set can reach a newer model's training data; a private one cannot be re-run by anyone else. The current case is the private one; these notes say so and take no position on the choice.

2 **Publication would not reproduce integers.** Even with a method published, judging again is a fresh draw, so a re-run means running the method and comparing distributions, not reproducing each score.

ON THE PROVENANCE OF THIS MATERIAL

The verdict record, cache and scoring plane described here come from ***FrontierScore***, a Digital One property, at frontierscore.ai; it is the provenance of this method, not its subject. Its repository is not public. The design description these notes draw on is its builders' own, and no part of it is checkable today, which is why its scale figures are left out. The public delayed feed is checkable, and each figure on this page about the live system was counted from it. Deliberately absent: prices, plans and tiers, availability, and deployment internals.

References

1 The Delegation Record, Paper 008, What these notes cannot do, and what would settle it.

2 The Project Handshake, Paper 004, What verification cannot see.

3 The Delegation Record, Paper 003, The field naming the model is filled in at the far end.

4 B. Atil et al., ***Non-Determinism of "Deterministic" LLM Settings***, arXiv:2408.04667v5, abstract and §1; task-accuracy experiments on five models, not judges. arxiv.org/abs/2408.04667, read 14 September 2026.

5 The Delegation Record, Paper 007, A switch attests what it decided, not that it decided well; its provenance note records that frontierscore.ai's public pages linked no methodology, read 12 September 2026.

6 The public delayed feed, `publicapi.frontierscore.ai/v1/scores` and `/v1/history`, responses saved at `generatedAt` 2026-09-14T14:28:00.505Z and 2026-09-14T14:28:00.963Z; the counts and correlations are this paper's, made from those saved responses, and a later fetch returns different values. publicapi.frontierscore.ai/v1/scores, read 14 September 2026.

7 frontierscore.ai homepage, visible text of the step headed "Score", saved 14 September 2026. frontierscore.ai (not linked), read 14 September 2026.

003

PAPER 003 · CONSTRUCTION

A measurement is a *signed claim*.

A probe fleet that signs its results can say which enrolled key signed a measurement and that the signed bytes have not changed since. It cannot say the measurement is true: **a signature binds bytes to a key, not to the world**. These notes set out what signing establishes in such a fleet — the encoding, the bundles, the keys — and **where the signed chain stops before it reaches a reader**.

PAPER · 003 PUBLISHED · 14 SEP 2026 READING · 8 MIN

LICENCE · CC BY 4.0

Canonical · digital1.foundation/articles/the-measured-endpoint/003-a-measurement-is-a-signed-claim.html

§1

A signature binds bytes to a key, not to the world

A probe fleet is distributed for a conditional reason: if an endpoint behaves differently depending on the network it is reached from, one measuring machine cannot see the difference. The condition is stated here, not found: these notes hold no measurement of endpoint behaviour by region. A fleet spread across networks, some of them run by parties other than the operator, is also an attack surface, and the design these notes draw on answers with signatures: each worker enrols an Ed25519 keypair and signs each result it submits, detached, over a canonical JSON encoding.

What that buys is exact and narrow. RFC 8032 gives Ed25519 deterministic signing, and signatures that "are not malleable due to the verification check that decoded S is smaller than l".[1] A valid signature establishes that the holder of an enrolled private key signed these bytes and that they have not changed since. It establishes nothing about whether the probe ran, whether the endpoint returned the transcript inside them, whether the timings are true, or when. A worker that fabricates a latency signs the fabrication as validly as a worker that measured one.

Signatures do change what agreement can achieve. Lamport, Shostak and Pease showed that "With unforgeable written messages, the problem is solvable for any number of generals and possible traitors".[2] That is agreement on one value. Colluding workers reporting the same false figure agree in exactly that sense, which is why Paper 004 treats consensus as agreement, not accuracy.

§2 Where the signing starts, and where it stops

The design names three signed objects: the test bundle a worker runs, the worker's own code, and each result it submits. What happens after submission happens in central planes and carries no worker signature. Judging of quality-scored output is done by the panel of Paper 001, which holds the rubric the worker does not,[3] and the aggregate is computed from what survived. Whether the operator signs those steps for its own use is not described; nothing in the two feed endpoints read on 14 September 2026 carries such a signature.

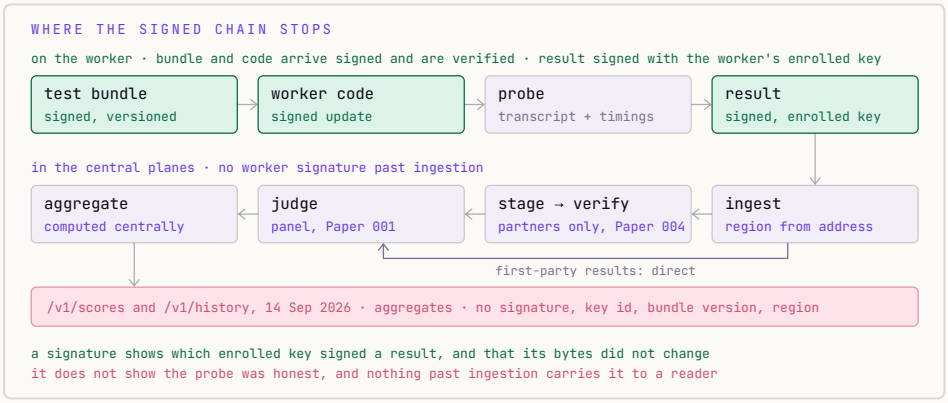

FIGURE 1 The signed part of the pipeline ends where the worker's part does. The bundle and the code arrive signed before they reach the worker and the result leaves under the worker's enrolled key; the probe between them is attested by nothing, and region, admission, judging and the aggregate rest on the operator's word.

The design's own description speaks of "a chain of custody from test definition to published score"; these notes cannot use the phrase. The chain stops at ingestion; to reach a reader, each aggregate would have to commit to the signed results it used, under keys a reader can obtain. A signature also cannot see what was never submitted or never promoted — in The Project Handshake's words, "the one attack it structurally cannot see is omission".[4]

OPEN PROBLEM · FOUND BY READING THE PUBLIC FEED FIELD BY FIELD, 14 SEPTEMBER 2026

Neither feed endpoint read for these notes lets a reader verify one signed result

Signing is meant to make a measurement checkable, and the check needs a result, its signature, the key and the encoding. The public feed publishes aggregates. This is `models[0]` of `/v1/scores` as fetched on 14 September 2026, every field shown:[5]

```
publicapi.frontierscore.ai/v1/scores · models[0] · generatedAt 2026-09-
    14T14:28:00.505Z
    { "provider": "openai", "modelId": "<model id>", "version": "<model id>",
      "accessPath": "api", "overall": <integer>, "quality": <one decimal>,
    "reasoning": <one decimal>,
      "speedTps": 294.3, "status": "healthy", "samples": 371,
      "asOf": "2026-09-14T13:26:32.133Z" }
    identity and score values withheld: the method that produced them is
    unpublished, see §5
    "version" is the model's version: it equals modelId outside the six anthropic
    series

    absent from each of the 18 records: signature · key id · worker id or tier
                                        bundle or test version · region
                                        commitment to the results aggregated
```

The 18 series in that snapshot, from five providers, carry the same eleven fields; `/v1/history` carries four per hourly point. Settling it would take published worker keys, a commitment per aggregate over the results it used, and their encoding — a decision for the operator, which these notes do not take. The score values are withheld above because they would read as a measurement of a named model: the method behind them is unpublished, and `overall` is a composite whose calculation the feed does not state.

§3

Canonical JSON is more than one encoding

The design's description says "Detached signatures over canonical JSON", which reads as a specification and is one name short of being one. A verifier has to rebuild the signed bytes exactly, and at least two documented schemes carry the name. RFC 8785, the JSON Canonicalization Scheme, says JSON numbers "MUST be expressible as IEEE 754 [IEEE754] double-precision values".[6] The dialect linked from The Update Framework specification, documented on the OLPC wiki, forbids what JCS admits: "Floating point numbers are not allowed in canonical JSON."[7]

That matters for this kind of record. The feed's records carry fractions — a `speedTps` of 294.3 in the record printed above, and scores given to one decimal place.[5] If signed results carry numbers like those, the scheme decides whether they are signed as numbers or must first be re-encoded as strings or scaled integers, and a verifier has to know which. The design names none, and that is a gap to state, not to fill.

What goes inside the signed bytes is a second choice. A result that does not bind its worker, bundle version, target, sequence and time window can be resubmitted into another window and still verify. The Project Handshake chooses its signed fields by enumerating the replay attacks and letting each one add exactly one field;[8] the design lists no signed-field set.

> THE REQUIREMENT, PRECISELY
>
> A signed result is checkable by someone other than its signer only when three things are published beside the key: the canonical encoding, the fields inside the signed bytes, and the checks a verifier applies. The design these notes draw on describes none of the three.

§4 A signed bundle stops substitution and little else

Tests reach workers as signed, versioned bundles, and worker code updates are signed and pulled from a pinned registry. Signing an artefact stops one attack: a substitute delivered in its place. The Update Framework specification, version 1.0.36, lists what an update system is meant to resist beyond that,[9] among them "Rollback attacks", "Indefinite freeze attacks" and "Mix-and-match attacks", and key compromise, which it answers with thresholds and separate signing roles.

Mapped onto a fleet — the mapping is these notes' reasoning, and nothing available to them shows the design implements TUF or part of it — each is a concrete failure. A worker held on an older bundle runs an older test and signs its results validly; that is caught only if results bind the bundle version and the centre refuses stale ones, and neither is described. A bundle at one version beside code at another may be a pair that never shipped together. If one online key signs bundles and code, it is one compromise away from signing for the fleet, which is why the specification keeps automated keys out of the roles clients trust for installation.[9] And pinning a registry pins where an update comes from, not which bytes it holds. A digest pins bytes.

Signing offers no secrecy. A bundle on a host a partner administers can be read by whoever administers it. The worker holds no rubric. If, as a test definition implies, the bundle carries the prompts, a partner host can read them, and a prompt set known outside the operator can be special-cased. The same exposure follows from publishing a prompt set so a method can be re-run; these notes do not resolve that tension.

§5

What a signature cannot carry, stated first

These notes describe an architecture. For this paper they publish no signed-field list, no encoding and no worker keys. The Delegation Record Paper 007 made a published method someone other than its author can re-run, or outcomes measured on a named corpus, the condition for calling a score sound, and recorded that frontierscore.ai's public pages, read 12 September 2026, link no methodology.[10] That condition is still unmet.

1. **A signature is as good as the key's custody.** On a host a partner administers, whoever administers it can read a software key and sign valid results with it. Binding a key to a measured environment is remote attestation, which the design does not describe.
2. **Past signatures outlive a compromise.** A result signed before a compromise was noticed still verifies. No revocation is described, and the record's account of what a revocation cannot recall applies.[11]

None of these is a reason not to sign results. They are the reasons to call a signed result what it is: a claim someone can be held to, not yet evidence someone else can check.

ON THE PROVENANCE OF THIS MATERIAL

The architecture described here comes from ***FrontierScore***, a Digital One property, and the design description these notes draw on was written by its builders. Its repository is not public, so that description — the signed objects, the encoding, the bundles and the registry, the staging of partner results — is their account and is not something a reader can check today. The public delayed feed at frontierscore.ai's API host is checkable while its seven-day window lasts, and the feed figures above were counted from snapshots of it taken on 14 September 2026. Deliberately absent: prices, plans and tiers, availability, and deployment internals.

References

1 S. Josefsson, I. Liusvaara, ***Edwards-Curve Digital Signature Algorithm (EdDSA)***, RFC 8032, January 2017, §8.2 (deterministic signatures) and §8.4 (non-malleability); Informational, IRTF CFRG. rfc-editor.org/rfc/rfc8032.txt, read 14 September 2026.

2 L. Lamport, R. Shostak, M. Pease, ***The Byzantine Generals Problem***, ACM Transactions on Programming Languages and Systems 4(3):382–401, July 1982, abstract. lamport.azurewebsites.net/pubs/byz.pdf, read 14 September 2026.

3 The Measured Endpoint, Paper 001 — ***A judge is a party to the score***.

4 The Project Handshake, Paper 004 — ***What verification cannot see***, on omission.

5 The public delayed feed, `publicapi.frontierscore.ai/v1/scores` (`generatedAt` 2026-09-14T14:28:00.505Z) and `/v1/history` (`generatedAt` 2026-09-14T14:28:00.963Z), both declaring `delayHours` 1 and `resolution` "1h". Every count here was made by these notes over the snapshots saved that day; a later fetch returns different values. An API, not linked; read 14 September 2026.

6 A. Rundgren, B. Jordan, S. Erdtman, ***JSON Canonicalization Scheme (JCS)***, RFC 8785, June 2020, §3.1; Independent Submission, Informational. rfc-editor.org/rfc/rfc8785.txt, read 14 September 2026.

7 OLPC wiki, ***Canonical JSON***, the dialect linked from The Update Framework specification §4.1; cited only to show the name covers more than one scheme. Wayback capture of 9 December 2025, web.archive.org/web/20251209150702/http://wiki.laptop.org/go/Canonical_JSON, read 14 September 2026.

8 The Project Handshake, Paper 003 — ***A signature that knows where it lives***.

9 ***The Update Framework Specification***, version 1.0.36, last modified 5 August 2026, §1.5 and §1.5.2. theupdateframework.github.io/specification/latest/, read 14 September 2026.

10 The Delegation Record, Paper 007 — ***A switch attests what it decided, not that it decided well***: the condition for a score's soundness, and the reading of frontierscore.ai's public pages of 12 September 2026.

11 The Project Handshake, Paper 006 — ***What a revocation cannot recall***: revocation as a claim about a set of actions rather than a state change.

004

PAPER 004 · LIMITS

Admission tests agreement, not *accuracy*.

Staging and consensus show that a promoted partner result **agreed with its peers and the operator's baseline, not that it was accurate**, and the public feed shows no individual result, promoted or rejected. A region derived from the source network shows where the traffic left as a mapping places it, not necessarily where the model ran. Per-provider circuit breakers keep a failing provider out of the probe budget apart from half-open trials, and **its worst hours can arrive as absent rows** the public history does not mark as not measured.

PAPER · 004 PUBLISHED · 14 SEP 2026 READING · 8 MIN
LICENCE · CC BY 4.0
Canonical · digital1.foundation/articles/the-measured-endpoint/004-admission-tests-agreement-not-accuracy.html

§1

Promotion tests agreement, not accuracy

The design these notes draw on, a probe fleet whose workers sign their results, has two trust tiers. First-party workers run on centrally held credentials and form the trusted baseline; their signed results go direct. Partner workers bring their own credentials and are untrusted until verified: each result they submit lands in staging and is promoted only after cross-worker consensus, a tolerance check against the baseline, and anomaly and fraud checks. The quorum, the tolerances and the fraud signals are not specified. The design also lets partner workers earn standing through verified contribution; what standing changes, and whether it relaxes any check, is not described. A standing earned by agreeing is spent by diverging: a partner honest until trusted and shaded after is caught only by checks that do not relax with reputation.

RFC 9334, read as an analogy — it concerns attesting a device's own state, and nothing here suggests hardware attestation — separates what the tiers blur: "Trust is a choice one makes about another system. Trustworthiness is a quality about the other system that can be used in making one's decision to trust it or not."[1] The tiers are trust in that sense.

Three consequences follow, as reasoning rather than measurement. Consensus among partners means something only if distinct keys are distinct operators: "if a single faulty entity can present multiple identities, it can control a substantial fraction of the system".[2] Douceur's setting has no central authority and this design has one at enrolment, so its defence is enrolment's check that two keys are two parties, which is not described. Colluding partners reporting the same shaded figure pass consensus by construction. And a tolerance against the baseline lets partner data confirm the operator more easily than contradict it: a degradation visible only from a network no first-party worker reaches looks, to that check, like an anomaly. That is the divergence a distributed fleet exists to find.

> PROMOTION IS NOT TRUTH
>
> A promoted result agreed with other results and with the operator's own baseline. That does not show the baseline was right, and no individual result, promoted or rejected, appears in the public feed, so the rate at which divergence was discarded cannot be seen from outside.

§2

A region is where the traffic left, as a mapping places it

The design takes a worker's region from its source network at ingestion, not from the worker's own report. That improves on asking the worker, and the design's description, which calls the region "verified", overstates it. Ingestion sees the address the traffic left from; turning that into a place takes a mapping the design does not describe — typically a geolocation dataset, some of it published by network operators about their own addresses. RFC 8805, the format for such feeds, tells a consumer to "only trust geolocation information for IP addresses or prefixes for which the publisher has been verified as administratively authoritative".[3] The self-report has moved from the worker to whoever publishes the address data.

Its accuracy has been measured, and is uneven. Poese and colleagues, in an editorial note whose header says it was not peer reviewed, on data from 2010 and 2011, wrote that "Geolocation databases can claim country-level accuracy, but certainly not city-level".[4] A cloud region usually names a metro. Weinberg and colleagues measured parties that choose where their traffic appears to come from: of the proxy servers they tested in 2018, "one-third of them are definitely not located in the advertised countries, and another third might not be".[5] Those were commercial proxy services, not a measurement fleet; the method-level point carries over to a partner worker on a network of its own choosing.

Two more distances are reasoning only: where traffic exited is not necessarily where the model ran, and it is not where users are. Active, latency-based geolocation is the kind of check that can bound a location and rule out a claimed one, and the design's description does not mention it. Neither `/v1/scores` nor `/v1/history`, as fetched on 14 September 2026, carries a region field; a third endpoint the operator's homepage reads was not fetched for these notes.[6]

§3

A provider's worst hours can arrive as missing hours

Provider failures — quota exhaustion, regional gating, bursts of server errors — are scheduling inputs in this design, with a circuit breaker per provider. In Fowler's description, once failures cross a threshold calls fail "without the protected call being made at all", and a half-open state makes "a real call as trial to see if the problem is fixed".[7] The design excludes open circuits before dividing the per-cycle probe budget, so the other providers' shares are computed without the failing one, apart from its half-open trials. That is a scheduling property, not a measurement one. Nor does it isolate the score: the judge panel of Paper 001 is drawn from the same providers, so an outage that opens one provider's circuit can also remove a panel member, and that touches the quality score of every provider's output, not only the failing one's.

The operator describes honoured backoff as capped at thirty minutes, whatever recovery time a provider advertises; that is its builders' description, and nothing a reader can fetch today shows it. What such a cap trades can be read against the specifications. A Retry-After sent with a 503 "indicates how long the service is expected to be unavailable to the client"[8] — advisory wording, so re-testing sooner breaks no rule, but where a provider asked for longer the trial lands inside that window. A 429 "indicates that the user has sent too many requests in a given amount of time":[9] a fact about the caller's quota, which a breaker keyed per provider would record as the provider's failure; the design does not say how it tells the two apart.

What a reader meets is simpler. While a circuit is open nothing is measured, so a provider's worst hours can arrive as absent hours rather than as low scores. In the public history fetched on 14 September 2026 — 18 series, 169 hourly timestamps, ten series complete — these notes counted the five openai series each missing the fourteen hours from 00:00 to 13:00 UTC on 11 September (two of them also missing 14:00 UTC on 7 September), and the three google series each missing ten or eleven of the twelve hours from 23:00 UTC on 7 September.[6] The feed gives no reason, and these notes attribute neither gap to a breaker. `/v1/scores` carries a current status ("healthy" or "elevated_latency" in that snapshot) but `/v1/history` carries none, so a missing hour is an absent

row and nothing marks it "not measured". Missing rows are not the only form a bad hour takes in that snapshot: 234 points carry a quality of 0, and the google gap is preceded, at 22:00 UTC on 7 September, by rows at quality 0 and reasoning 0 in all three series. Nothing in either endpoint distinguishes a judged 0 from an hour in which no verdict was produced. An average over the week therefore silently drops some bad hours and may count others as zeros, and a reader cannot tell which.

```
/v1/history · generatedAt 2026-09-14T14:28:00.963Z · hourly ts missing, UTC
openai  5 series   2026-09-11T00:00–13:00, in all five
                   2026-09-07T14:00, in two of them
google  3 series   2026-09-07T23:00 to 2026-09-08T10:00, ten or eleven hours each
the other ten series have all 169 hourly points · model identities withheld
```

§4

What this cannot establish, stated before anyone asks

These notes describe an architecture. They publish no rubric, no prompt set, no panel composition and no test suite — and, for this paper, no admission thresholds. The Delegation Record Paper 007 made a published method someone other than its author can re-run, or outcomes measured on a named corpus, the condition for calling a score sound, and recorded that frontierscore.ai's public pages, read 12 September 2026, link no methodology.[10] That condition is still unmet. A reader can check the reasoning, the cited specifications and the shape of the feed, and cannot check that the fleet does what is described.

1. **The tier the admission checks skip is the operator's own.** First-party results form the baseline without passing the checks partner results pass.
2. **Nothing here measures the fleet.** The design's own size and throughput figures are left out because none can be checked from anything public.

ON THE PROVENANCE OF THIS MATERIAL

The architecture described here comes from ***FrontierScore***, a Digital One property, and the design description these notes draw on was written by its builders. Its repository is not public, so that description — the two trust tiers (first-party and partner), the admission steps, the region taken at ingestion, the per-provider circuit breakers, the thirty-minute cap — is their account and is not something a reader can check today. The public delayed feed at frontierscore.ai's API host is checkable while its seven-day window lasts, and the feed figures above were counted from snapshots of it taken on 14 September 2026. Deliberately absent: prices, plans and tiers, availability, and deployment internals.

References

1 H. Birkholz, D. Thaler, M. Richardson, N. Smith, W. Pan, ***Remote ATtestation procedureS (RATS) Architecture***, RFC 9334, January 2023, §1; Informational. Used as an analogy. rfc-editor.org/rfc/rfc9334, read 14 September 2026.

2 J. R. Douceur, ***The Sybil Attack***, IPTPS 2002, abstract and §1. microsoft.com/en-us/research/wp-content/uploads/2002/01/IPTPS2002.pdf, read 14 September 2026.

3 E. Kline et al., ***A Format for Self-Published IP Geolocation Feeds***, RFC 8805, August 2020, §3.2; Independent Submission, Informational. rfc-editor.org/rfc/rfc8805.txt, read 14 September 2026.

4 I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, B. Gueye, ***IP Geolocation Databases: Unreliable?***, ACM SIGCOMM Computer Communication Review 41(2), April 2011, abstract. An editorial note, not peer reviewed per its own header; data from 2010–2011. Text read from the ORBi repository copy, orbi.uliege.be/bitstream/2268/107073/1/p53-v41n2h2-uhligA.pdf, read 14 September 2026.

5 Z. Weinberg, S. Cho, N. Christin, V. Sekar, P. Gill, ***How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation***, IMC '18, abstract and §8. Author-hosted copy, contrib.andrew.cmu.edu/~nicolasc/publications/Weinberg-IMC18.pdf, read 14 September 2026.

6 The public delayed feed, `publicapi.frontierscore.ai/v1/scores` (`generatedAt` 2026-09-14T14:28:00.505Z) and `/v1/history` (`generatedAt` 2026-09-14T14:28:00.963Z), both declaring `delayHours` 1 and `resolution` "1h". Every count here was made by these notes over the snapshots saved that day; a later fetch returns different values. The history window is seven days, so these gaps are no longer in the live feed after 18 September 2026; the missing timestamps are printed in §3. An API, not linked; read 14 September 2026.

7 M. Fowler, ***CircuitBreaker***, bliki, 6 March 2014; the pattern is credited there to M. Nygard. martinfowler.com/bliki/CircuitBreaker.html, read 14 September 2026.

8 R. Fielding, M. Nottingham, J. Reschke (eds), ***HTTP Semantics***, RFC 9110 (STD 97), June 2022, §10.2.3 (Retry-After) and §15.6.4 (503). rfc-editor.org/rfc/rfc9110.txt, read 14 September 2026.

9 M. Nottingham, R. Fielding, ***Additional HTTP Status Codes***, RFC 6585, April 2012, §4 (429 Too Many Requests). rfc-editor.org/rfc/rfc6585.txt, read 14 September 2026.

10 The Delegation Record, Paper 007 — ***A switch attests what it decided, not that it decided well***: the condition for a score's soundness, and the reading of frontierscore.ai's public pages of 12 September 2026.

END

THE MEASURED ENDPOINT · COLOPHON

Colophon

A

The record

This document is generated from the published pages, which are the record. Where this file and a page disagree, the page is right and this file is stale. Each source below is named with the SHA-256 of the file this document was built from, so a reader can check that what they are holding is what was published; the generator fetches each live page as it builds and reports any that has moved since.

INDEX digital1.foundation/articles/the-measured-endpoint/

001 digital1.foundation/articles/the-measured-endpoint/001-a-judge-is-a-party-to-the-score.html
sha256 7b2cde015656bbd1dff9d91e353ae087555cd336e71b7c3459a34243c35c8892

002 digital1.foundation/articles/the-measured-endpoint/002-a-verdict-can-be-kept-not-re-run.html
sha256 88c209dd5887ad08e8f7104b5d27a015f34223d2544bf8b7fcfaef0dd6180c7f

003 digital1.foundation/articles/the-measured-endpoint/003-a-measurement-is-a-signed-claim.html
sha256 db8176acc580f093ba4d13d590a6c480e1eb19a00e2aeaa23e3e96e74dd712df

004 digital1.foundation/articles/the-measured-endpoint/004-admission-tests-agreement-not-accuracy.html
sha256 8b82596fc8eec4c2c58de04c321b3af8d056688471055ebf60efc103c9c67193

B

Licence

C

How it is set

In the six faces the site itself carries and embeds: Fraunces 300 and 600 for titles and the emphasised word, Manrope 400 and 500 for the text, JetBrains Mono 400 for the ledger's own labels, code and running matter. Those six are Latin subsets and do not carry every character the papers use — → is absent from them. Nothing is drawn in their place and quietly dropped from the text: each is set in a named system face. Every

character in every paper was extracted back out of this file and checked before it was written, so the whole document copies, searches and cites as it reads.

Figures are the papers' own inline drawings, re-toned from the site's night ground to this page by a one-to-one swap of the closed house palette, so green still means it holds and rose still means it breaks — and the two are separated in weight of grey as well as in hue, so a photocopy keeps the distinction. The rail down the left of each page is the ledger from the Foundation's homepage; a filled dot marks each numbered section, and the § number sits in the gutter, outside the measure.

The sheet is A4 as the page asks for it, which Chromium rounds to 594.96 × 841.92 pt when it writes the file — 0.1 mm narrower and 0.03 mm taller than the 595.28 × 841.89 pt of the standard. No printer will show it; a document that prints its own hashes should say it anyway.

---

digital1.foundation
The Measured Endpoint