nature portfolio

Peer Review File

Lu⁺ optical frequency references with accuracy verified at the 19th digit

Corresponding Author: Dr Kyle Arnold

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Please see the reviewer attachment.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
Review Comments for Authors

This manuscript demonstrates a landmark achievement in optical timekeeping. The authors set a new benchmark in systematic uncertainty for an optical clock, reaching 1.1 \times 10^{-19} for their 176Lu+ system. The central and most significant result is the direct validation of this accuracy via a frequency comparison—enabled by correlation spectroscopy—between two independent clocks, showing agreement at the level of -2.4 \pm 5.8 \times 10^{-19}. This represents the first demonstration of a clock's performance below the 10^{-18} level validated by a same-species comparison.

The result is a major advance for the field. While the use of 176Lu+ and techniques such as hyperfine averaging and correlation spectroscopy builds on prior work by this group and others, the achievement of validated 10^{-19}-level performance is unprecedented, representing a significant technical milestone. Its importance lies in providing the first conclusive, self-consistent validation at this level, thereby meeting the stringent 2 \times 10^{-18} threshold set by the CCTF for redefining the SI second.

That said, the strength of this work lies in its profound technical depth and its role as a definitive benchmark for the atomic physics and metrology community. While the result is monumental, its core impact is vertical—deeply advancing a specific field—rather than horizontally broad, which journals such as Nature often emphasize. I believe this work certainly deserves publication as the next step beyond recent PRL papers by Aeppli et al. (Sr, 8.1 \times 10^{-19}, PRL 2024), Marshall et al. (Al+, 5.5 \times 10^{-19}, PRL 2025), and Zhang et al. (Ca+, 4.4 \times 10^{-19}, PRL 2026), and would be most appropriately published in a specialized, high-impact physics journal such as Physical Review Letters.

The experimental approach is sound and the data quality is very high. The presentation of the systematic uncertainty budget is clear and detailed. However, two aspects of the methodology require strengthening:
(1) Radial Heating Rate Anisotropy:
Figure 6A reports the heating rate for only one of the two radial motional modes. In linear Paul traps, the two radial directions are not necessarily equivalent due to potential asymmetries in electrode geometry or the surrounding electrical environment.

Heating rates can differ significantly between these modes because of variations in coupling to electric field noise, patch potentials, or dielectric losses.
Request: Please provide the measured heating rate for the second, orthogonal radial mode. Both radial heating rates should be included, for example in the supplementary information.
(2) Blackbody Radiation (BBR) Shift Assessment:
The BBR shift is a dominant systematic uncertainty in high-accuracy optical clocks. Currently, the Methods describe its evaluation based solely on temperature measurements of the two “end blocks.” This approach is likely insufficient and may underestimate the true uncertainty, because the temperature at the end blocks need not represent the temperature of the electrodes at the ion’s position—the relevant parameter for the BBR shift. In both the Lu-1 and Lu-2 traps, the RF and DC electrodes experience ohmic heating from applied voltages and likely follow different thermal pathways, meaning the electrode temperature could be substantially higher than that of the more massive end blocks.
Request: The authors should provide a more rigorous, quantitative assessment of the BBR shift. This should include:
(i) Enhanced thermal analysis, which could be achieved via direct temperature measurements closer to the trap electrodes, or through a well-validated finite-element thermal model that incorporates RF/DC ohmic heating and the detailed trap geometry. Alternatively, a holistic evaluation might involve measuring the temperature and emissivity of each material inside the vacuum chamber, along with the solid angle subtended at the ion.
(ii) An updated uncertainty estimate for the BBR shift based on a realistic upper-bound for the electrode temperature, rather than relying solely on the end-block temperature. An explicit discussion of the likely thermal gradient between electrodes and end blocks is necessary, and its contribution should be included quantitatively in the overall uncertainty analysis.

Minor Comments:
(1) The “Hyperfine-Averaged Hyper-Ramsey” sequence used to suppress probe-induced shifts is clever. To assist other groups in implementing similar protocols, could the supplementary material provide further details on the exact pulse sequence—for example, a diagram or a list of phases—along with the derived robustness conditions against common-mode pulse-area errors?
(2) The long-term stability of the \(\alpha_{QZ}\) coefficient over three years (Fig. 3) is remarkable and is a key pillar of the low uncertainty. Were these measurements taken under identical trap conditions (e.g., ion position, RF power)? Could the authors comment on any observed sensitivity of \(\alpha_{QZ}\) to trap alignment or other parameters that future experiments should control?
(3) The BBR shift is the main area where the uncertainty analysis could be made more rigorous. Additionally, could the authors clarify whether any potential correlations in environmental noise (e.g., magnetic field fluctuations) between the two independent traps were considered in the uncertainty of the frequency comparison?
(4) Please update the references to include formal journal citations for entries currently cited as arXiv preprints that have since been published.

Referee #4

(Remarks to the Author)
Review for manuscript entitled “Optical clocks with accuracy validated at the 19th
digit” by K.J. Arnold et al.

The authors report a frequency comparison between two independent $^{176}Lu^{+}$ single-ion clocks using correlation spectroscopy with a shared clock laser. This technique enables an instability of 4.8×10^{-16} τ^{-}1/2. The systematic uncertainty is evaluated at 1×10^{-19} based on measured field sensitivities and coefficients, while the total measurement uncertainty of ± 5.3×10^{-19} is currently limited by statistical instability.

1) The reported relative frequency agreement is dominated by a statistical uncertainty of $\pm 5.3 \times 10^{-19}$, which exceeds the quoted systematic uncertainty. While the result is consistent with agreement at the 10^{-19} level, the measurement does not statistically resolve agreement at that level. Given the broad readership of Nature, the authors should clearly clarify this distinction and ensure that claims of 10^{-19} performance are framed as inferred or consistency-based rather than experimentally demonstrated.

2) “Here we report two 176Lu+ single-ion optical clocks, each with fractional frequency uncertainty near 1 ×10−19, a five-fold improvement over the highest accuracy reported to date.”
The use of correlated spectroscopy with a shared interrogation laser removes laser noise by design, which is appropriate for probing atomic agreement and common-mode rejection, but does not capture the performance of two independent clocks, for which the clock laser is an integral component. The authors should clarify that the reported agreement reflects atomic reference consistency under shared-laser interrogation, rather than standalone clock performance.

3) This reviewer finds the following text misleading: "We demonstrate interrogation times up to ten seconds and a comparison instability of 4.8 ×10−16 (τ/s)−1/2, nearly a factor of 3 improvement over previous work [16]. Comparing optical references via correlation spectroscopy has √2 higher instability than comparing independent clocks, and a factor of two higher instability than an individual clock, for the same interrogation time. It follows that these 176Lu+ systems are currently capable of~2.4 ×10−16 (τ/s)−1/2 clock instability if combined with state-of-the-art laser stabilization technology [28]. This is an improvement over the lowest instability demonstrated in an ionic frequency standard [12], and would enable instability below 10−18 in less than one day."

The final inference is problematic. The projected single-clock instability relies on assumptions that are not met experimentally: namely, that a single clock can be interrogated for the same duration as enabled by correlation spectroscopy (which is unlikely given that trap lifetimes of typically of the order of ~1 s due to micromotion and magnetic field noise effects), and that state-of-the-art cryogenic laser stabilization (absent from the present system) is available. As written, this risks conflating demonstrated comparison performance with hypothetical single-clock capability. As a result, this statement should be revised or removed.

While the authors claim and improvement of a factor of 3 as compared to previous demonstrations, the comparison is relative to the author's previous measurements [16], and does include a comparison to the state-of-the-art from other groups. This self-comparison needs to be clarified in the text, or the claims should be made against the previous work of the broader community:
• Aeppli, Alexander, et al. "Atomic clock frequency ratios with fractional uncertainty <= 3.2 x 10-18." arXiv preprint arXiv:2512.21428 (2025).
• Kim, May E., et al. "Improved interspecies optical clock comparisons through differential spectroscopy." Nature Physics 19.1 (2023): 25-29.
• Clements, Ethan R., et al. "Lifetime-limited interrogation of two independent Al+ 27 clocks using correlation spectroscopy." Physical review letters 125.24 (2020): 243602.

4) The language used in much of the text, but especially in the Experiment Set up and Results sections, seems to be targeting an audience very familiar with atomic clock operation. The authors rely heavily on undefined technical terms such as "hyper Ramsey spectroscopy", "correlation spectroscopy", "Lamb-Dicke" parameters, "hyperfine averaging" etc. As such, in its current form is not written in such a way as to be accessible to the broad readership of Nature. At the least, both sections should be rewritten, replacing jargon with accurate but simpler language explaining clock operation, notable differences or similarity to other groups and systems, and a clear explanation of how the Lu+ clock operation enables advantages in performance as compared to previous reported measurements.
• For example, correlation spectroscopy is used to probe both ion samples simultaneously with a shared laser, which helps to suppress laser noise and enables longer interrogation times (Ramsey times).
• Hyperfine averaging: averaging over magnetically sensitive transitions helps to suppresses the first-order Zeeman shifts on the clock transition frequency.
• These types of explanations would help orient readers prior to Eq. (1).

The work is technically strong and addresses an important question in optical clock performance. However, for Nature's broad readership, the current framing overstates the level of experimental validation and relies heavily on specialist language. With clearer distinction between inferred systematic control and statistically resolved performance, explicit discussion of the implications of correlated spectroscopy (a method developed elsewhere), and more accessible description of the experimental protocol, it is unclear whether to gains in performance represent sufficient novelty for publication in Nature. As written, and given the current results in comparison with the state-of-the-art, this manuscript seems more suitable for PRX or Nature Physics.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
\documentclass{article}
\usepackage{graphicx}
\usepackage{amssymb}
\usepackage{enumitem}
\setlist[enumerate]{label*=\arabic*.}

\title{Referees report on 2025-12-32495A}
\author{}

\begin{document}

\maketitle
This review attempts to walk a somewhat fine line. On the one hand, we continue to believe that the experimental work reported here is both high-quality and of significant and broad enough interest to merit publication in Nature. This work marks several significant steps forward for atomic clocks, at a time when the redefinition of the second amplifies the importance of atomic clock progress. We discuss this further below.

On the other hand, this revision has added a substantial focus on the authors' opinions about the state of atomic clock research and the shortcomings of the SI redefinition process. We have significant issues with these opinions and their inclusion in this report. The recommended changes conflate the clock evaluation with the more rigorous process of validation; they would likely not resolve the issues with comparison repeatability cited by the authors; the underlying thesis is applied in a logically inconsistent way; and the proposed criteria are not fully met by the experimental work presented here. We will also discuss each of these points in more detail below.

In sum, we do not believe that the manuscript should be published in its present form, but we hope the authors will revise and continue to seek publication in Nature. We believe that a version of this article which emphasizes the impact of the experimental work will satisfy other referees' concerns about broad interest, without centering the authors' opinions at the expense of their remarkable experimental achievement.

Other points are included at the end of the review, including concerns about the unmeasured secular-motion heating rate, where we note that published works and preprints have reported up to a factor of 5 anisotropy in macroscopic, symmetric traps.
\\\textbf{Regarding impact and interest}

Both of the other referees expressed the opinion that this result is not of sufficiently broad interest to serve Nature's readership. We disagree for several reasons, which we will summarize here.

First, there is the matter of multiple-record-setting performance. This is the most accurate clock ever constructed, by a factor of more than four. In addition to that evaluation, this work also reports the most precise clock comparison to date (and one of the most precise measurements of anything ever). We believe these records alone carry a certain amount of interest to physical scientists, and papers reporting record advancements in clock performance have been reported in Nature - e.g. Bloom et al., Nature 506, pages 71–75 (2014) reporting the JILA strontium lattice clock.

Second, this is the first comprehensive accuracy evaluation of a lutetium clock. This is an important milestone, since lutetium promises significant advances in accuracy compared to better-established clocks. In this sense, we find it analogous to initial evaluations of other types of clock previously published in Nature, e.g. King et al., Nature 611, pages 43–47 (2022) reporting the first comprehensive evaluation of a highly-charged-ion clock, as well as the many Nature reports of incremental progress towards a thorium nuclear clock. We note that this is especially important at a time when the CCTF is establishing the short-list of atomic clock candidates for the upcoming redefinition of the second; this work makes a strong argument for lutetium's relevance to that process. Perhaps this could be emphasized in a future revision.

Third, the combination of a systematic uncertainty evaluation with a same-species comparison is a significant advancement beyond the single-clock evaluations that the other referees compare to this work. While we disagree (see below) with the authors that same-species comparisons should be required before any clock evaluation can be published, it is indisputable that a same-species comparison makes a clock evaluation stronger; it bounds many types of possible experimental error, should make interspecies comparisons involving the clocks more reliable, and demonstrates clock performance in a more realistic scenario than an evaluation alone. We note that the only analogous work of which we are aware was also published in Nature, as McGrew et al, Nature 564, pages 87–90 (2018) reporting on the NIST ytterbium lattice clocks.

We also hope that the reorganization of the paper, with many technical details moved to the Methods section or supplemental material, and with important references and explanations added to many important points, will make it more accessible to non-specialists.
\\\textbf{Issues with the revised manuscript}

Having made the case that the scientific merits of this work rise to the high level expected from a Nature publication, we move to reasons we do not believe it should be published in its current form. We continue to have issues with the authors' discussion of ``validation'', as well as their new criteria for ``acceptable'' atomic clock evaluations. We do not find the authors have adequately addressed our concerns on this front; instead they have devoted significant new emphasis to their interpretation of ``validation'', which we find flawed, as well as to their criticism of atomic clock research more broadly.

Before getting into the specifics of this debate, we note that a peer-reviewed research report is not the appropriate forum for such opinion. This would conflate peer review of the research with peer review of the opinion. Large parts of the new ``discussion'' section are essentially unrelated to the research work being reported here, instead being part of the authors' criticism of the state of the field. Rather than inside a high-profile research report, we believe the community and the authors would be better served if these criticisms were published separately in a journal like Metrologia. That would be a more fitting venue for public debate; and the criteria could be presented as an aspirational standard for clock groups including the authors, rather than undermined by being published alongside a comparison which does not fully meet them.

There are a few separate issues here which we will break down.
\\\textbf{Regarding ``validation'' as a term of art}

The authors' new definition of ``validation'' is significantly less rigorous than the community consensus included in the SI redefinition roadmap, mandatory criterion I.2 (N Dimarcq et al., Metrologia 61 012001 (2024), ref. 2 of the revised manuscript). The core element lacking from the authors' definition is reproduction by independent experimental groups. ``Validating'' an atomic clock requires reproduction by different institutes, which is an essential check against experimenter error, subconscious bias, challenges from long-distance time transfer or clock transportation and unidentified shifts. This is fundamentally not something that can or should be done by a single institute.

For example, the authors cite the NIST Yb optical lattice clock (ref. 15 of the revised manuscript) as the only other atomic clock which meets their definition of a ``validated'' accuracy budget. We point them to the appendices of arXiv:2512.21428 (ref. 8 of the revised manuscript). Here, this same group discusses how, because of errors in atomic structure calculations, the systematic shifts applied in ref. 15 were too small by 3E-18 -- more than double the total systematic uncertainty evaluated in ref. 15. Clearly, the presence of a same-species comparison with high precision did not guarantee that the accuracy budget was correct.

Additionally, the degree of independence of the two clocks matters when determining whether the evaluation can be considered ``validated''. As an example of the kind of unidentified shifts which are not eliminated in this comparison, we note that spectral impurities in a clock laser can cause -- and have caused, in optical lattice clocks at least -- significant systematic shifts, which would cancel in common mode in a comparison like the one performed here. Requiring that comparisons be between different institutions establishes an objective baseline for what can be considered ``independent'', which otherwise is highly subjective (as one can clearly see from our review and the authors' response).

This contention over ``validation'' is in part a semantic issue, since -- as we discussed in our previous report -- ``validation'' already has a specific definition in the context of next-generation clock uncertainty estimates. The authors have a legitimate claim that their uncertainty evaluation is better supported by evidence than others which do not include self-comparisons. A different, less loaded term would avoid this conflation of the two definitions -- perhaps ``confirmation'', ``verification'', or ``corroboration''. However, both because of the common institution and because of the reduced level of independence of the two clocks, we continue to hold that this comparison doesn't reach the level of a fully ``validated'' frequency reference.
\\\textbf{Does the lutetium clock meet the authors' criteria?}

Semantics aside, we believe there are issues with the application of these criteria to the present work.

First, as applied to lutetium the criteria are logically inconsistent. By insisting that clock accuracy budgets are only ``acceptable'' insofar as they have been tested with same-species comparisons, the authors imply that rigorous evaluation via secondary measurements is insufficient or ``unacceptable'' -- even where no specific reason exists to doubt such an evaluation. The argument is that to be trustworthy, these evaluations require additional information from a same-species comparison to corroborate the standard uncertainty evaluation. At core, this is an epistemological stance advocating skepticism towards a specific category of scientific results. For this stance to be meaningful, it must be applied uniformly. If only applied where convenient to the authors, it ceases to be rigorous.

The same-species comparison reported here provides no additional information to corroborate the evaluations of the blackbody shift or second-order Zeeman shift, since those are both common-mode between the clocks. Therefore, those aspects of the clock evaluation don't satisfy this epistemological criterion and should not be considered ``acceptable''.

In response to this point from our initial review, the authors claim that ``both of these properties have been rigorously assessed''.
This is true; the scientific work in this manuscript is of high quality. However, systematic effect evaluations by many other high-quality research groups around the world also feature ``rigorously assessed'' properties with equally high quality. Insofar as the authors wish to discount those other accuracy evaluations because they haven't been tested in same-species comparisons, the exact same logic must apply to these shifts on lutetium. To be ``acceptable'', they must be corroborated by additional information from a same-species comparison. Applying this standard to an overall clock evaluation but doing so in a way that does not corroborate the two largest shifts is logically inconsistent.

We note that a same-species comparison at significantly different temperatures could, in principle, uncover errors in calculated or measured atomic properties; in that case, the incorrect BBR shift corrections applied for the two different temperatures would result in a non-unity ratio.

In this light, no clock evaluation can be ``rigorous'' under the authors' criteria until a same-species comparison has been made between two apparatus at significantly different temperatures and with significantly different magnetic fields. That this would be difficult is not sufficient reason to make an exception just for lutetium or just for the BBR shift; it's evidence that ``validating'' a clock without relying on rigorous characterization of effects through secondary measurements is, in fact, difficult.

We emphasize that we don't consider this a reason to doubt the evaluation of the BBR shift in this work. Rather, this undermines the authors' asserted criteria for clock evaluations and their criticism of standard practices in the clock field.
\\Second, the role of the same-species comparison in making an accuracy evaluation more rigorous is to constrain shifts due to unknown effects and due to experimenter error. A self-comparison can constrain these only insofar as they are not common-mode between the two clocks. However, the independence of these two systems is limited in ways that significantly reduce their ability to constrain unknown effects, even beyond the temperature and magnetic field servos. Shared environments, control software, data analysis pipelines, apparatus design, and other operating conditions are all avenues that unknown effects could come in through. To our knowledge, the majority of errors found in previous uncertainty evaluations would not have been caught by a same-species comparison similar to the one performed here; for example, the incorrect atomic structure coefficients or clock laser spectral impurities discussed above would both cancel in common mode.

One of the reasons the SI redefinition roadmap demands different institutions make comparisons is to control for these unknown effects. So-called ``dark'' uncertainty is unfortunately common in high-precision interspecies clock comparisons, both within comparison campaigns and between measurements of the same ratios (cf. references in section 6C of Fortier et al., Optica 13 143-163 (2026), as well as Aeppli et al., arXiv:2512.21428 (2026)). Same-species comparisons involving independent systems can help resolve these issues precisely by constraining unknown shifts affecting one or more clocks.
\\Third, this work doesn't meet the authors' own ``minimum criteria for a rigorous assessment''. For one, they are making a ``performance claim'' at the 1e-19 level but only do a comparison at the 5e-19 level. In the authors' language, the ``claimed accuracy cannot be better than the measurement precision'' for the same-species comparison. If the authors were reviewing

a different group's clock accuracy evaluation, one assumes that they would implement this standard by refusing to accept an accuracy claim with precision beyond the relevant self-comparison. The same should be true here - if this performance claim is ``unacceptable'', it should be moderated to match this standard or the standard should be rethought.

For another case, the new criteria require ``a stress test on any systematic that is more than an order of magnitude larger than the accuracy claim''. As far as we can tell, the authors have not done sufficient ``stress tests'' on the two clocks presented here. Again, this would require changing the BBR environment significantly enough to evaluate the BBR shift. While the authors argue the differential BBR shift is small for a reasonable range of laboratory temperatures, what needs to be epistemologically confirmed and ``stress tested'' isn't the differential shift between the two clocks; it's the total BBR shift from 0K to room temperature, which is more than an order of magnitude larger than the claimed accuracy.

The measurements using different fields on Lu-1 (presented in Fig. 3) would satisfy this requirement for Lu-1 on the Zeeman shift, although it's not clear whether that applies to Lu-2 as well -- that depends on whether the authors intend this to be a requirement only on a species-by-species basis, or separately for each clock evaluated.
\\\textbf{Are these new criteria for ``acceptable'' clock comparisons a good idea in the first place?}

The authors would like to add same-species comparisons to mandatory condition I.1, when they are already present in a more restrictive form in I.2. Aside from this being insufficient to demonstrate ``validation'' as defined in criterion I.2, we disagree that a same-species comparison should be required in order to report clock accuracy evaluations. We believe this stems from a different understanding of what an accuracy evaluation is or should be. As interpreted in nearly every other published optical clock result to date, the accuracy evaluation is nothing more nor less than a collection of measurements which rigorously estimate shifts and uncertainties to an atomic clock frequency. In other words, the authors' criterion i.

In essence: an accuracy evaluation is not a ``hypothesis'', as the authors put it in the revised introduction. It is a specific type of reported research result.

An accuracy budget does not amount to a guarantee of the variation in frequency ratio measurements from a ``true'' value -- that includes the significant additional experimental effort which goes into making a high-precision comparison, including frequency transfer, phase-noise compensation, operation of optical frequency combs for inter-species comparisons, relative geopotential measurements between clocks, and data analysis. All of these are opportunities for error which cannot be appropriately included in any single clock's accuracy evaluation.

Adopting the authors' criteria would amount to invalidation of every previously published accuracy evaluation except one, which features identified flaws. Additionally, it would push institutions to redirect resources towards replicating identical apparatus rather than developing new clocks or advancing clock performance, and it would exclude groups without the resources to build two identical clocks.

Moreover, any additional utility from this definition is already captured by the existence of mandatory criterion I.2 -- as part of the SI redefinition process, clock error budgets must be validated at 5E-18 by multi-institute clock comparisons. Beyond the redefinition, in the case of the optical second, the CCTF Working Group on Primary and Secondary Frequency Standards would continue to stringently evaluate individual frequency standards at the relevant level of precision, just as they do now for both microwave and optical standards at the E-16 level relevant to TAI.

We agree with the authors that same-species comparisons have additional value beyond the accuracy evaluation. They represent a check against many - but not all - potential significant errors in clock construction and evaluation. They also demonstrate the performance of a clock in a more direct way. Hence, again, why we support publication of the underlying results here in Nature rather than a less prestigious subject-matter-specific journal. However, this utility is additional to the accuracy evaluation, not inherent to it.
\\\textbf{Magnetic field correlation analysis}

We agree that the magnitude of the magnetic field noise will not be a significant issue, and the added details make this claim more convincing. Our comments and questions have been addressed. However, the newly added Fig. S2d requires some additional details. First, there is no metric provided for judging the magnetic field noise as `uncorrelated.' One could claim a small anti-correlation based on visual inspection alone — what is the correlation coefficient? What does the thick black line represent? Second, the quantum projection noise, which is independent between Lu-1 and Lu-2, can dominate the correlation of the magnetic field noise. For the experimental conditions in Fig. S3a, the averaging time would need to be around 100 seconds to remove the projection noise component.
\\\textbf{SODS and heating rate measurements}

We were disappointed to learn that the authors had not measured the secular motional energy in one of the three secular modes for both clocks. The authors are correct about the distance scaling of anomalous heating, and that heating rates typically scale smoothly with frequency, but both significantly anisotropic heating and anomalous frequency scaling have been reported in other macroscopic traps. For example, Dawel et al (arXiv:2509.22525, 2025) observed a factor of 5 larger heating rate along one radial direction than the other, using a blade trap generally similar to that used in this work and with a difference in secular frequencies of less than 0.5\% between modes. Lindvall et al (PR Applied 24, 044082, 2025) observed heating rates that changed by a factor of 5 over a measurement campaign, and depended on trap secular frequency beyond a smooth scaling. Leopold et al (Rev. Sci. Instrum. 90, 073201, 2019) observed a factor of 2 difference in heating rates for a cryogenic blade trap with a $\sim$10\% difference in mode frequencies.

We note that while the trap used in this work demonstrates mechanical symmetry, there is not a matching electrical symmetry - according to ref. 50 of the revised manuscript, one pair of rod electrodes provides the RF pseudopotential and the other pair is held at DC. This implies different impedances to ground and different electrical sources connected to the pairs, which is a natural way anisotropic technical noise could affect the ion.

This is an area where the same-species comparison is valuable - realistic sources of anisotropic heating are likely technical in nature rather than inherent to the trap design, so they're unlikely to be correlated between the two systems. However, this only constrains the effect at the 5.7E-19 level, rather than the 1E-19 overall accuracy. Perhaps the authors can derive a more stringent constraint from the dephasing measurements, or measure the heating rate of the inaccessible mode using parametric mode coupling as demonstrated in Gorman et al (Phys. Rev. A 89, 062332, 2014). However, it's not obvious to us, given these results from a brief literature search, that potential anisotropy can be accounted for simply by increasing the uncertainty on the unmeasured mode from $\sim$5\% to $\sim$10\%.

\\\textbf{First-order Doppler shift}

We agree with the authors' discussion of charging and discharging from UV photoelectrons. This is an unrealistic possibility when applied to the Lutetium case. The point of bringing up the NIST group's measurement was not to suggest a particular source of the FODS, but rather to make the point that large first-order Doppler shifts can arise in ion trap systems and be difficult to detect without counterpropagating beams. The authors point out an issue with the photoelectron hypothesis from that experiment - if they are correct, that would only further emphasize that first-order Doppler shifts can arise without an obvious mechanism, and should be constrained by measurement.

Photoelectrons aside, the authors have three responses to our concerns about the potential first-order Doppler shift, which we will summarize here for clarity of discussion.
\begin{enumerate}
\item The self-comparison constrains any FODS, because the odds that it would be identical and synchronized between clocks are vanishingly small
\item The authors dispute the likelihood of the specific possible FODS scenarios we discussed in our initial report.
\item Constraining the FODS with new measurements using counterpropagating beams would take too long to average down, and ``incorporating a purely hypothetical FODS into the uncertainty budget lacks scientific utility''
\end{enumerate}

We will address each of these points in turn. Regarding the first point - we agree that the odds of an undetected FODS are small, and it is reasonable to constrain it at the level of the self-comparison. However, this comparison is five times less precise than the claimed systematic uncertainty, so this point is only partially relevant. This is related to the contradiction discussed above between the authors' performance claims and their epistemological stance on the necessity of self-comparisons.

Regarding the second point - we initially note that all of these proposed mechanisms were intended to illustrate the difficulty of constraining an FODS from first principles, rather than detailed estimates. In line with the authors' epistemological arguments, first-principles arguments such as these are relatively unsatisfying. A few specific notes:
\begin{itemize}
\item We agree that scattering from the ion is negligible. To our knowledge, the MW power used is not given in the manuscript; at single-digit milliwatts, we agree that is also unlikely to cause substantial differential heating across the trap.
\item The optical powers used for interrogation were similarly not given in the manuscript. If the figure of 10 μJ per clock cycle includes laser cooling and readout as well as clock interrogation, then we agree the total power is likely too small to cause a substantial FODS. We appreciate the authors' provision of additional details.
\item Regarding Lorentz forces, the concern is not oscillating currents, which as the authors point out average to zero for currents at different frequencies, but rather DC forces from power supply currents and ground return currents, which may create a different electric-field environment when the laser or microwaves are off vs on. This is also likely too small to matter but again requires microscopic knowledge of the experimental apparatus, e.g. where do ground currents flow / how well have ground loop currents through the vacuum chamber been eliminated, etc.
\item The timescale argument is only partially convincing. For example, suppose there is a mechanical vibration on the milliseconds scale from a shutter closing. To average 30 pm/s over the seconds-long probe time, the scale of the vibration at the ion or at one of the clock laser beam-path mirrors would have to be 30 nm/s, which still does not seem unreasonable if the shutter is close by and depending on the spring tension in the relevant mirror mounts.
\item Relatedly, we note that ion clock experiments have observed first-order Doppler shifts related to the transient response of path-length servos to mechanical-shutter-induced vibrations. We believe this is already constrained by the AOM phase chirp measurement, but mention it to point out that the timescale question can be quite complicated - in this case, the transient response extends the timescale of the perturbation significantly beyond the timescale of the mechanical vibration.
\end{itemize}

Regarding the authors' third point - we fail to see a meaningful difference between incorporating a ``purely hypothetical'' error source into an accuracy budget just because it isn't experimentally constrained, and rejecting rigorously evaluated uncertainty budgets that don't come with a self-comparison to reject other purely hypothetical effects.

That aside - while it may be unsatisfying, we acknowledge that the standard practice in the field is to omit the FODS from uncertainty budgets unless there is specific reason to include it. We believe this leaves unanswered the question of whether large FODSes of unknown origin, similar to that observed by the NIST group referenced earlier, may also be present on other ion clocks. We believe this work would be better-supported had the authors conducted their campaign with counterpropagating clock beams. However, we acknowledge that they are reasonable at the level of standard practice in the field, especially when combined with the additional information provided in their response but not included in the manuscript, and don't require any additional changes to the manuscript on this front. We appreciate the inclusion of bounds on some FODS sources that can be estimated from the data the authors did collect.

That said, there is a certain tension with the authors' arguments about what constitutes an ``acceptable'' or ``unacceptable'' accuracy claim, where they discount rigorous evaluation and first-principles arguments as unverified without a self-comparison. Are the authors truly satisfied with this evaluation of the first-order Doppler shift, and if so what differentiates it from other clocks' ``rigorous evaluation'' of systematic effects which the authors consider ``unacceptable?'' The measurement of the first-order Doppler shift using counterpropagating clock beams would constrain otherwise un-measureable effects in the same way that a self-comparison would constrain some unknown differential effects; however, the authors decided not to make this measurement, while simultaneously claiming that accuracy budgets are not acceptable with uncertainty below the stability of a self-comparison.

If the authors want to present their clock as not just a significant experimental achievement, but as a model for scientific rigor and example of a framework that other atomic clocks should follow, they must also address the places where their work is in tension with that framework. Similarly to the lack of ``stress tests'' discussed earlier in our review, the fact that many potential sources of FODS are constrained only by first-principles argument rather than measurement should be addressed, or else the framework and epistemological arguments should be modified to be consistent with the experiment.
\\\textbf{Small comments}

Minor typo on page 2 of supplemental material: "We take this power to be half the total rf drive power assuming equal dissipation at each end of the trap \textbf{and}, i.e. all the power is dissipated in the ceramic."

Regarding the BBR shift: We appreciate the additional detail included in the new supplemental material. The authors have adequately supported their estimates of the BBR shift.

Similarly, we find that the questions about the magnetic field servo have been addressed.

\end{document}

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)

Referee #4

(Remarks to the Author)
Reviewer #4 (follow-up, April 21, 2026) for manuscript entitled, “Optical clocks with accuracy validated at the 19th digit” by K.J. Arnold et al

I appreciate the authors’ efforts to revise the manuscript and their detailed responses. The work remains technically strong and addresses an important problem in optical frequency metrology. However, several of the central concerns raised in the initial review are only partially addressed, and in some cases the revisions remain difficult to assess based on the response provided.

Accessibility and clarity for a broad readership
The authors indicate that the manuscript has been substantially revised for accessibility, including moving technical material to the Methods and adding explanatory text. However, the response does not identify specific changes or provide representative examples demonstrating how key concepts have been clarified.

Given the importance of accessibility for Nature’s readership, it is important that central ideas such as correlation spectroscopy, hyperfine averaging, and the structure of the interrogation sequence are introduced with clear conceptual explanations prior to their quantitative use. Based on comparison with the original manuscript, the revisions appear to consist primarily of minor rewording and relocation of technical material, rather than the addition of conceptual explanations needed for accessibility to a broad readership. Ultimately, the revisions are difficult to verify and appear limited in scope.

Demonstrated, inferred, and projected performance
The distinction between experimentally demonstrated performance and inferred or projected capability remains insufficiently explicit. While the authors have moderated some language, projections of single-clock instability based on correlation

spectroscopy measurements rely on assumptions (e.g., availability of state-of-the-art laser stabilization and comparable interrogation times) that are not realized in the present work. These projections should be clearly identified as conditional and separated from the primary experimental results.

Interpretation of performance under shared-laser interrogation
The manuscript would benefit from clearer framing of what aspect of “clock performance” is being probed. Correlation spectroscopy provides a powerful measure of atomic reference agreement under shared-laser interrogation; however, for a broad readership, this is not equivalent to the performance of a fully independent clock system.

A more explicit distinction between 1) atomic reference consistency under common-mode rejection, and 2) standalone clock performance would help avoid potential misinterpretation of the results.

Framing of “validation” and statistical resolution
The revised manuscript introduces a more explicit definition of “validation,” which is helpful. However, the central issue remains that the measured agreement is limited by statistical uncertainty at the 5 x 10-19 level and therefore does not resolve agreement at the 10-19 level. It is important that this distinction is consistently and unambiguously reflected in the manuscript, so that claims of performance at the 10-19 level are clearly framed as inferred or consistency-based rather than experimentally resolved.

In summary, while the manuscript has been improved, further clarification is needed to ensure:
• accessibility to a broad readership,
• clear separation of demonstrated and inferred performance, and
• unambiguous framing of what aspect of clock performance is being measured.

Addressing these points would significantly strengthen the manuscript and clarify its suitability for Nature. As written, the technical specificity to those very familiar with optical metrology and optical atomic clocks makes of the manuscript more suitable for PRX.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have significantly revised the manuscript, including its title. The revised manuscript presents the context in a more natural way and thus describes the standing of the Lu+ frequency standard more clearly. The verification of its performance is now presented in a more self-contained and self-consistent manner, so that readers in the community can appreciate the significance of the work.
We largely agree that an uncertainty budget can be seen as a testable prediction about same-species comparisons between co-located clocks, since this removes much of the potential error in multi-species comparisons which cannot be clearly attributed to a specific clock.
We also agree with the authors’ new framing of the same-species comparison as a rigorous test of the accuracy of the uncertainty evaluation, rather than a component of it.
We still disagree with the authors about the desirability of counterpropagating probe beams, especially when the claimed systematic uncertainty is well below the statistical uncertainty of the self-comparison. While there is more to say about specific hypothetical scenarios, such a debate could continue indefinitely. Instead, at this point we will simply restate our central argument: 30 pm/second (a 1E-19 clock shift) is an extremely small amount of motion, and it’s difficult to prove the absence of cycle-synchronized uV-level changes in trap potentials or sub-nm changes in trap dimensions or position via logic alone. Counterpropagating probe beams comes at no cost to clock stability or accuracy once operational and can easily suppress a first-order Doppler shift well below the statistical uncertainty of a clock comparison. In general, this seems in line with the authors’ philosophy that clock claims should be tested with experimental data where possible.
All that being said, we reiterate for the editors that the authors’ treatment of this effect is in line with the current standard practice in the field and should not be a barrier to publication. We only include this brief restatement as the authors’ “rebuttal” indicated an interest in continuing this debate for the peer review file.
We appreciate that the authors went back and measured the heating rates in the other radial direction, especially since they indeed found a significant difference in the heating rate for one of the traps. While it is true that this difference was well below the statistical uncertainty of the comparison, we note that it was larger than the claimed differential systematic uncertainty. The additional radial heating measurement completes the evaluation.
Overall, we find that the improved manuscript is suitable for publication in Nature.

Minor comments:
Line 96: it would be useful to state what is being averaged (e.g., environmental field fluctuations and laser intensity fluctuations).
Line 114: typo, missing "an" in "an additional phase shift"
Line 303: typo, missing space in "30 s in"
Line 454: typo, missing space in "corresponds to"

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks to the Author)
I appreciate the substantial effort the authors have made in revising the manuscript. The presentation has improved in several respects, and several of the more speculative claims have been moderated. However, I remain unconvinced that the manuscript meets the threshold for publication in Nature.

My principal concerns are as follows:

Accessibility to a broad scientific readership: Despite the revisions, I do not believe the manuscript is sufficiently accessible for Nature. The paper remains highly specialized and assumes considerable prior knowledge of optical clock operation, systematic uncertainty evaluation, and frequency ratio measurements. Much of the motivation, experimental work, and results and impact are presented at a level of specialization that I believe will be difficult to follow even for researchers in neighboring fields such as photonics or laser science. Given Nature's broad readership, substantially greater effort is required to explain why this experiment matters beyond the optical clock community and what broader scientific question is being addressed.
Significance of the reported performance: The manuscript reports a comparison between two Lu+ optical clocks with a systematic uncertainty approaching 1×10^(-19). While this represents an impressive technical achievement, the reported measurement uncertainty of approximately 5.7×10^(-19) limits the extent to which this systematic performance can presently be validated experimentally.

It is therefore unclear that the advance over the existing state of the art is sufficiently compelling for Nature. As one benchmark, the 2018 Nature demonstration comparing two independent ytterbium optical lattice clocks reported a systematic uncertainty of 1.4×10^(-18), a measurement instability of 3.2×10^(-19), and reproducibility from blinded comparisons at the 10^(-18) level. While the present work substantially reduces the evaluated systematic uncertainty, the experimental comparison itself is performed with a measurement uncertainty of approximately 5.7×10^(-19). It is therefore less clear that the work establishes a correspondingly new level of experimentally demonstrated clock agreement.

I would also like to acknowledge an important distinction between the present work and the 2018 ytterbium lattice clock comparison. The Lu+ system is based on a single trapped ion, making this the highest-accuracy comparison reported for a single-ion optical clock to date. Compared with neutral-atom optical lattice clocks, which interrogate approximately 104 atoms simultaneously, single-ion systems inherently exhibit lower signal-to-noise ratios and higher quantum projection noise, resulting in reduced measurement stability. From this perspective, the work represents a significant technical milestone for single-ion optical clocks. My comments above concern whether the level of experimentally demonstrated advance, when viewed in the broader context of the field, reaches the editorial threshold expected for publication in Nature.

I think the most compelling demonstrations at this level of uncertainty are likely to be comparisons between fully independent clock systems (interspecies or same species) operating below the 10^(-18)level. Such experiments provide a more stringent validation of clock accuracy because they minimize the possibility that shared local oscillators, laboratory environments, experimental infrastructure, or correlated systematic effects contribute common-mode biases. Although this is not a requirement for publication, it illustrates why I find the overall significance of the present advance somewhat less compelling than the manuscript suggests.

In summary, the work represents a significant technical accomplishment and will likely be of considerable interest to the precision metrology community. However, in my opinion, the combination of limited accessibility for a broad readership and the degree of experimentally demonstrated advance relative to previous landmark optical-clock comparisons means that the manuscript does not currently meet the editorial threshold expected for publication in Nature.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Citation reference numbers in both referee comments and author's response refer to the bibliography of the first version of the submitted manuscript. Any numbered references which are discussed in this reply are listed at the end of our response to avoid any confusion.

GENERAL COMMENTS TO REVIEWERS AND EDITORS

We thank the reviewers for their time and thorough evaluation of our manuscript. The reviewers raise several valuable points, particularly regarding the blackbody radiation (BBR) shift, which we have carefully addressed in the revised text. Reviewers 1 and 2 speculate on numerous hypothetical mechanisms for first order Doppler shifts for which it is difficult to provide a definitive technical response without further context and specified parameters. We have made every effort to address these comments comprehensively, providing technical bounds for these effects wherever physically applicable.

In accordance with the reviewers' feedback, we have revised the manuscript to ensure it is highly accessible to Nature's broad, interdisciplinary readership. We have streamlined the main text by migrating highly specialized experimental details to the Methods and Supplemental Material sections. Additionally, we have incorporated clearer, more generalized explanations of our core methodologies, including hyperfine-averaging, hyper-Ramsey spectroscopy, and correlation spectroscopy, to better highlight their broad utility. Finally, we have expanded our discussion on the definition of 'validation' within the context of the SI Second Roadmap, providing a framework to compare our results against recent advancements in the field.

In the following sections, we provide general comments addressing the two main areas of critique: our meaning of 'validation' relative to the Roadmap, and the blackbody radiation shift assessment. We also include a brief discussion on magnetic field correlation and its implications for clock instability to collectively address related issues raised by multiple reviewers. All remaining queries are addressed in the subsequent point-by-point response.

Roadmap and Validation

From the beginning of our Lu^+ research we set out to build two clocks guided by the view that any performance claim (evaluated uncertainty) is merely a hypothesis that must be tested experimentally by same-species comparison, i.e. a unit-ratio. We stipulate here (and in the revised manuscript) three criteria for the assessment of optical frequency standards (OFS) we consider the minimum for acceptance of a performance claim:

i a clear set of experimental measurements of atomic properties and environmental factors determining the systematic uncertainty budget,

ii a same-species comparison to demonstrate performance and reproducibility of the OFS, where the claimed performance cannot be better than the measurement precision,

iii a stress test on any systematic that is more than an order of magnitude larger than the accuracy claim, i.e. a frequency comparison in which the systematic is deliberately shifted to observe the predicted shift in the clock frequency.

In the present work for Lu^+ we are not claiming 'validation' beyond this definition. Specifically, we claim that (i) all known systematics have been rigorously evaluated to $\sim 1 \times 10^{-19}$, and that (ii), performance is 'validated' at the level of $\sim 5 \times 10^{-19}$ by demonstrated agreement limited at the statistical precision.

The only shift significantly larger than the measurement precision is the quadratic Zeeman shift, which, consistent with our criterion (iii), has been explicitly stress tested by comparisons made over a large magnetic field range. Moreover, it has been measured in three different years with the latter two measurements done after extensive lab renovations and reorganization of the lab.

The revised manuscript makes clear our definition of validation and in the discussion we highlight why this distinction is important.

BBR shift

The BBR shift assessment is based on bounding the temperature between the ambient temperature within the enclosed space in which the chamber sits and the maximum temperature the ion could reasonably see. The test trap is the former Lu-1 from [16] which was removed from the existing chamber, rebuilt due to the heating issue reported in [16], and reinserted into an identical chamber. This test trap has exactly the same construction with exactly the same mounting as the traps used for the comparison. Ceramics used in building a trap are purchased in quantity and typically 10 pcs so we had stock on hand to test from these original batches. The PT100 sensors attached to the test trap are as close as practical to the center of the trap, and that is on the mounting blocks at each end of the trap. They cannot be attached to the ceramic as it is embedded in the mounting block. It cannot be attached to the electrodes, nor should it be because that is not where the power is dissipated.

We bound the temperature between the ambient temperature and the maximum we can reasonably estimate. Previously we just took this maximum to be no more than twice the measured temperature rise at the measurement point. The reviewers have correctly pointed out that electrodes may be at a higher temperature than measured at the end block. We now include an analysis of the power dissipation in the ceramic to estimate the maximum electrode temperature increase relative to the copper end block where measurements were made. The technical details of the temperature evaluation have been moved to Supplemental Material. Now that we have a reasonable estimate for the maximum electrode temperature, we take the midpoint temperature between min and max temperature, with 1σ capturing the min and max, to be consistent and allow for normal statistical variance beyond the evaluated min and max. However, we note that to even match the $\sim 1.0 \times 10^{-19}$ uncertainty contribution from the polarizability corresponds to a maximum temperature bound approximately 8 degrees above the ambient. Evaluation at this level does not require the extensive simulations of an effective blackbody environment and the influence of gradients, emissivities of materials, etc., as is needed in other systems.

Magnetic field noise

We provided general comments to clarify several points raised across the reviewers' comments regarding magnetic field noise and the implications for clock stability. First, regarding the sensitivity to magnetic field noise. All states used are $m = 0$ which have no first order Zeeman shift. The $^3D_1$ $|F\rangle$ states have quadratic Zeeman shifts $\delta\nu_{QZ,F} = \alpha_F B^2$ where $\alpha_F \approx [2320, -140, -2180]$ Hz/mT^2 for $F = [6, 7, 8]$. Every single HA-HR interrogation sequence averages over the three states which leaves only the remaining Quadratic Zeeman shift for the HA average, which is $\delta\nu_z = \alpha_z B^2$ with $\alpha_z \approx -4.8$ Hz/mT^2 as reported in the paper. Note the dominant sensitivity of the component states comes via the α_6 and α_8 which are roughly equal and opposite and so the symmetric nature of the averaging (8-7-6-7-8) further suppresses sensitivity to linear changes in the field over a single interrogation duration. In the revised supplemental we now present data for the magnetic corrections by the field servo for both chambers during a 37 hour continuous run. Note that the magnetic field corrections are uncorrelated between the chambers and that without any stabilization whatsoever the magnetic field would have stayed within a ± 60 nT range corresponding to $\pm 1 \times 10^{-19}$ relative change in $\delta\nu_z$ at $B = 0.1$ mT. This is just to highlight that the magnetic field servo is largely done for convenience and not in any way essential for $< 10^{-19}$ control of the quadratic Zeeman shift, which could equally be corrected point-by-point from the interleaved linear Zeeman measurement to trivially obtain much better than 10^{-20} uncertainty. We discuss the possibility of field servo error in the new supplemental, but these must be negligible by any measure. As already noted in methods, for the stabilized field the shift averages down as $\sim 2 \times 10^{-19} \cdot (\tau/\mathrm{s})^{-1/2}$ so is exceedingly small after minutes of averaging, much less days.

Atomic coherence times are, however, limited by the short timescale magnetic field noise within a single HA-HR interrogation (5-10 seconds). The sensitivity to magnetic field fluctuations is given by the derivative of the quadratic Zeeman shift $\frac{d\delta\nu_{QZ,F}}{dB} = 2\alpha_F B_0$, which when operating at $B = 0.1$mT, is $\sim \pm 400$ Hz/mT for $F = 6$ and 8. For reference, this is over four orders of magnitude lower than linear Zeeman sensitivity of alkaline earth clock transitions e.g. Sr^+, Ca^+, Ba^+. As shown in methods, from independent measurements on field sensitive states, we observed $\sim$7 nT flicker noise on the magnetic field in each chamber. We do not expect magnetic dephasing to be correlated as the chambers are separated by 1.2 m, the applied dc magnetic fields are different spatial orientations in the lab frame, and <0.5m from each chamber are different server racks filled with power supplies and electronics. The observed contrast loss in correlation spectroscopy (Fig. 2a) is consistent with a simulated servo for **uncorrelated** flicker noise in each chamber. We note the microwave transitions [ref 18] have the same sensitivities ($\sim$400 Hz/mT at 0.1 mT) and Ramsey decoherence times on these transitions for the individual systems are consistent with the measured magnetic flicker noise. For increased B_0, we observe both correlation spectroscopy and microwave clock spectroscopy coherence times decrease proportionally, consistent with an increase in sensitivity ($2\alpha_F B_0$) but no change in the ambient noise amplitude. Insofar as the magnetic noise is indeed uncorrelated, as the evidence suggests, it directly follows that the current systems would be capable of the same interrogation time operated as clocks if a suitably stable laser were available. We believe this is an important point: optical clock interrogation time, and hence clock instability, can be limited by either optical or atomic coherence. For an atomic system with large environment sensitivities or finite lifetimes this can be a significant limitation, we are simply pointing out the implicit evidence-based capability of the current Lu^+ system. And in the conclusion, we point out the future potential, as this magnetic noise limitation is rather a trivial technical problem to overcome. We reasonably anticipate the next iteration of the system will be able to demonstrate atomic optical coherence times on the minute timescale.

POINT-BY-POINT RESPONSES

Referees' comments (blue), Authors response (normal text)

Response to Reviewers #1 and #2

1. Overview

The manuscript reports on two optical frequency references with systematic uncertainties of about 1×10^{-19} and demonstrates a frequency difference measurement at 5×10^{-19}, limited by statistical uncertainty. This is the first full evaluation of the Lu^+ clocks, after years of efforts by the authors, and establishes Lu^+ as an excellent frequency reference. This work sets two world records, for clock accuracy and clock comparison precision. Moreover, each is a significant improvement the systematic uncertainty is five times lower than the existing record, and the comparison uncertainty is twice as low as the previous best. The combination of these two achievements is more than the sum of its parts the comparison helps to justify or build confidence in the accuracy, a step beyond previous record-setting clock evaluations. At a time when the global scientific community is undertaking a major effort to redefine the SI second, we believe this manuscript is appropriate for the wide readership of Nature.

The manuscript is largely valid, with careful evaluations of many potential sources of systematic error. The systematic evaluation as it stands currently is incomplete, as we discuss below; but we are confident the authors can address our concerns in revision, either with additional measurements or further analysis. The manuscript is presented efficiently, but contains sufficient information for the expert reader to fully evaluate the results. Statistics are evaluated appropriately using standard tools in the field.

In sum, we believe this manuscript merits eventual publication in Nature. However, before publication we have several concerns that must be addressed. Two significant systematic shifts need additional attention the first order Doppler shift and the evaluation of the effective blackbody temperature. See below for detailed discussion of these, as well as for comments and concerns about the rest of the manuscript.

Before the specific elements, however we wish to raise a general point which recurs across the manuscript. This concerns the discussion of “validation” and “independent” clocks. We agree with the authors that same-species comparisons are essential to progress towards the optical second. The work presented here does represent the most accurate same-species comparison to date and is a significant step for ion clocks and for optical clocks generally; the inclusion of a same-species comparison together with the uncertainty budget is what brings this work up to a higher level.

We thank the reviewers for their recommendation and acknowledging the importance of the same-species clock comparison in combination with the systematic evaluation. As detailed in the general comments, we believe we have addressed the main concerns of the BBR evaluation. We respectfully disagree with much of the reviewers comments on first order Doppler shift for the reasons given in point-by-point responses, but now include a first order Doppler shift discussion in the revised manuscript insofar as is justifiable.

That said, we find the authors’ discussion of the same-species comparison as an independent validation of the uncertainty budget to be slightly overstated. These two clocks are independent in that they have separate vacuum chambers, but their systematic effect evaluations still overlap to a significant degree. They are servoed to the same magnetic field and environmental temperature; share the same laser and RF sources; are operated with synchronized clock cycles (meaning they share laser noise, environmental noise, etc); are operated under a single differential servo algorithm; and are run, designed, and analyzed by the same scientists. The comparison validates the evaluation of differential systematic shifts, such as AC Zeeman shifts, second-order Doppler shifts, etc. However, the two largest systematic shifts quadratic DC Zeeman and blackbody radiation are essentially common-mode. Similarly, the comparison does not bound unknown common-mode systematic effects. We find no issues with how these effects are evaluated the authors have done careful and thorough work in this and other papers to measure these effects. The point is that the same-species comparison does not make these parts of the evaluation any more “valid” than a careful evaluation with a single clock, as errors with their evaluation would cancel out of the comparison.

We maintain that the shared operational parameters listed do not diminish the independence or validity of the comparison. Specifically, the uncertainties of the dc quadratic Zeeman and BBR shifts are dominated by uncertainty in α_z and the polarizability, which are atomic properties and inherently common-mode to any same-species comparison, regardless of whether the clocks are in the same room or on different continents. Both of these properties have been rigorously assessed. The quadratic Zeeman shift is specifically measured by comparison with large magnetic fields differences and has been measured three times, in three different years, with major changes to lab organization subsequent to the first measurement. Similarly, the polarizability has been evaluated through multiple consistent measurements, including the dc polarizability, the dynamic polarizability at several wavelengths, and key contributing matrix elements. While it is true all of these measurements have so far been done by a single team, namely ours, this risk also applies for most other systems where all clocks of a given species may rely on a single group’s measurement of the relevant atomic property. More importantly, because Lu^+ possesses exceptionally low sensitivity, the scale for common-mode errors is orders of magnitude smaller than in other widely used species.

For Lu^+ temperature has very little influence on the assessment owing to the extreme insensitivity and not because the two temperatures appear common-mode. To illustrate: an 8 K difference in temperature corresponds to a shift of only $\pm 1.0\times 10^{-19}$ about the mean BBR shift of the two systems. This temperature tolerance is sufficiently broad it would capture the ambient differences between practically any two laboratories globally. This contrasts sharply with systems where large BBR shifts require extensive simulations, parameter optimizations and accurate calibrated temperature measurements, such as for Lindvall et al, PRApplied 24 044082 (2025) noted by the reviewer. Lindvall et al claim a 51 mK accuracy for their Sr^+ clock. A comparison of two such systems on the same optics table as we have done, would not constitute a common mode thermal environment.

To be clear, we are not questioning the thoroughness of the temperature evaluation in Lindvall et al. (as but one example), but rather the suggestion that it does not warrant empirical confirmation through a comparison, which runs counter to basic scientific principles. The further suggestion that a comparison made within the one lab has no value beyond the uncertainty assessment of a single clock because it does not bound an unknown systematic that is inexplicably common-mode, is an unreasonably dismissive standard.

The authors should address these limits, whereas the manuscript in its current form could be interpreted to claim that the comparison completely validates the full uncertainty evaluation. For example, the authors could clarify what is meant by "independent" in the context of the two Lu^+ clocks by adding a sentence in the "discussion" section, perhaps in the context of the existing discussion on line 207 of the degree to which perturbations are uncorrelated; or alternatively use more specific language when describing how the two clocks are independent on the first page; etc. Additionally, the claim on lines 75-76 of "strong evidence that all systematic shifts above the 5×10^{-19} level have been accounted for" should be moderated this is only true for differential systematic effects between the two clocks.

The closing statements of the intro, and the discussion, have been expanded to clarify these points.

We also point out that the term 'validation' is more rigorous in the roadmap defined from the international metrology community (I.2 category, ref. 2). At least three experiments (refs. 6, 7, 15) have shown unity frequency measurements below the 5×10^{-18} criterion. Although these works have similar or greater independence compared to this work, the community does not consider them as 'validation' since the comparisons were performed within a single laboratory. Given this apparent consensus on the term 'validation' within the metrology community, and given that the authors also take such consensus seriously, the claim of validation should not be misleading. While the authors discuss this more stringent interpretation of the term 'validation' in the Discussion section

Insofar as the roadmap (ref. 2) represents the consensus, we disagree that any assessment (I.1) should be considered valid without the minimum demonstration of reproducibility that is a comparison of two separate systems, i.e. validation.

... ,given this difference in definition and the fact that only some effects are addressed by the comparison, we question whether the term "validated" is justified in the article's title.

We now rigorously define what we mean by this and have suggested how the Roadmap criteria could be strengthened.

In the following, we first give comments on the main text and then go through each systematic shift. The two systematic shifts we mentioned earlier are discussed in the 'First order Doppler shift' and 'BBR assessment' section.

2. Main text

On line 21, 'difference' instead of "agreement"?

changed to "...demonstrating agreement in the measured relative frequency difference..."

On lines 66-72 the authors state that the potential instability of a single lutetium clock, with an ideal noiseless clock laser, of 2.4×10^{-16} at one second is an improvement over the lowest instability demonstrated by a trapped-ion clock. This claim is somewhat misleading a better apples-to-apples comparison would be with the manuscript's reference 27, where a correlation spectroscopy stability of 2.8×10^{-16} was achieved. The same analysis would imply a laser-noise-free limit of 1.4×10^{-16} at one second.

We have moderated the statement and removed any claims about 'best' stability. Still, we believe it is important to point out the implied instability our system is capable of as inferred from the differential atomic decoherence time, which is limited only by technical sources in the present system.

On lines 75-76, while the statistical uncertainty of the comparison is 5×10^{-19} and the measurement is consistent with zero, it's not obvious that 5×10^{-19} is the right level to claim justification of the differential uncertainty evaluation. The $1 - \sigma$ confidence interval is $(-7.8, +3) \times 10^{-19}$, so errors in the evaluation of differential systematic effects are only excluded with 67% confidence below 7.8×10^{-19}.

There are three numbers given: the measured outcome, the statistical uncertainty, and the systematic uncertainty. The systematic uncertainty concerns the best we could hope to achieve, the statistical uncertainty the measured precision. The outcome of any similar measurement campaign would be governed by a random variable with standard deviation determined by the rms sum of the statistical and systematic uncertainty. Here, for all practical purposes, that is the statistical uncertainty. A measurement outcome would be compared to that total uncertainty. The null hypothesis is that the references agree. The measured outcome is within 1σ so we have no reason to reject the null hypothesis. If it was out by 2σ, we would reject the null hypothesis at the 95% confidence level. Adding the single outcome value to the uncertainty to render a bias and then taking the largest of the two resulting values as a bound is not a standard, or correct, statistical interpretation.

Regarding Fig. 1 is AOM 1a used to cancel common-mode path length fluctuations between the ions and the shared clock laser source?

No, correlation spectroscopy is not sensitive to clock laser noise or any common-mode path fluctuations, only differential laser phase fluctuations. The differential path is stabilized from the point where the beam is split on the PBS up to the retroreflecting mirror reference points using AOM 2b. AOM 1b does nothing but provide a fixed frequency offset. AOM 1a and 2a switch the optical powers and apply the phase shifts and frequency steps required for the HA-HR sequence. The frequencies of $f_{1a} + f_{1b} \approx f_{2a} + f_{2b}$ so that the clock laser is resonant in both chambers, but $f_{1a} \neq f_{2a}$ and $f_{1b} \neq f_{2b}$ so that the detected reference beat for the differential path stabilization, $|2(f_{1b} - f_{2b})|$, is well resolved from all rf frequencies f_{1a}, f_{1b}, f_{2a}, f_{2b} and free of signal cross-contamination.

Also, delivering the maser signal accurately at 3×10^{-15} is also not so trivial. Could you describe how the setup is designed?

The maser resides in the same climate controlled lab and is distributed over relatively short (< 20 m) cable lengths from one low-noise distribution amplifier (Spectra Dynamics HPDA-15MRi-C). A Spectra Dynamics HROG-10 offset generator is used for steering the maser reference signal. The maser instability is characterized in [50] (absolute frequency measurement) and the phase stability of the distribution network is discussed in [18] (precision microwave frequency measurements). Both references have been added to the 'rf synthesis' section.

The comparison measurement is anyway not sensitive to maser accuracy, and after further consideration, we realize that for single clock operation the accuracy of the hyperfine average is also not limited by the accuracy of microwaves as long as the clock frequency is coherently synthesized from the fields used in interrogation as $\nu_0 = f_L + \frac{1}{3}(2f_1 + f_2)$. The discussion in the methods has been updated to reflect the correct considerations for single clock operation.

Regarding Tab. 1, the background gas shift is 0. Is this because of the collision detection algorithm described in Methods?

No. As discussed in our paper ref 19, the phase of the integral used to determine the shift is very sensitive to changes in the molecular potential (see Fig. 11 in ref 19) such that the imaginary component is more suitably bounded by the absolute value. We would consider any bias from a limited range of calculated potentials to be merely fortuitous.

On lines 82-83, the authors mention two central methods, hyperfine-average and hyper-Ramsey without any citation or explanation. It will be useful to have citations and brief descriptions for the broad scope of readers (what is averaging and not averaging in HA, probe Stark is canceled in HR). Besides, the pulse sequence seems to be slightly different from [16].

We thank the reviewer for pointing this out, it was an oversight to not have citations at that point in the manuscript. This section has been significantly rewritten.

Regarding stark shift suppression by HR, the sequence used in [16] and the time reversed HR sequence used in this work are functionally equivalent, in the absence of thermal effects. However when including ion heating, they are perturbed in the opposite way, and by averaging the two in a hybrid signal, ion heating effects on the HR stark shift suppression can be effectively eliminated. We had a figure and discussion prepared in the methods section on this but it was removed before submission, as the technique is not used here (standard HR suppression was already sufficient). It has been reintroduced into methods as the difference between the two sequences may be of interest to the reader, as well as the availability of this simple technique for further suppressing thermal effects.

On line 84, $\sigma_{z,i}$ is not defined.

Removed $\sigma_{z,i}$. Now include a citation at the point where the parity is introduced for the reader not familiar with parity and correlation spectroscopy.

On line 85 - We couldn't find a definition of $f_{6,7,8}$ anywhere in the text or figures. If ν is used for representing atomic frequency and f for the field frequency, $\nu_{6,7,8}$ can be used with the definition. On line 87, is $f^{(k)} = f_L + (2f_1 + f_2)$ or $f^{(k)} = f^{(k)} + 1(2f_1 + f_2)$? The current form gives $\delta f = 0$ always.

This section has been substantially rewritten and everything is now defined.

Regarding Tab. 2, there are several points that should be addressed. The systematic uncertainty of the reference 15 should be 1.4 instead of 6. The 6 contribution is reported as the gravitational redshift from the Terrestrial Time surface, which can be neglected in local measurement like this work. A recent work from VTT MIKES could also be included in this table [T. Lindvall et al., Phys. Rev. Applied 24, 044082 (2025)].

We thank the reviewer for both points, the incorrect uncertainty on reference [15] was an oversight, and we had not been aware of the recent Sr^+ result. Another recent result for Sr(USTC) is now also included.

The authors include their previous work with a total systematic uncertainty. We could not find the explanation of both systematic uncertainty and the comparison uncertainty. In the reference 16, the authors "refrain from making any accuracy claim" because of the lack of a BBR assessment.

The paragraph from ref [16] in full is:

An accuracy claim for an individual system would require an assessment of the BBR shift. Presumably, the use of simulations and thermal cameras (2, 15) could equally be used here, although it is unclear how to quantify such an assessment for our system. A crude analysis as given in Materials and Methods would bound the temperature to 35(10) C, which, due to the low BBR shift for the $^1S_0 \leftrightarrow ^3 D_1$ *transition, corresponds to a BBR shift of just* $-1.56(27) \times 10^{-18}$ *and gives a total uncertainty for both systems of* $< 8\times10^{-19}$. *However, although it is done elsewhere (2, 37), we refrain from making any accuracy claims without a same-species comparison to the appropriate level. Here, stability limits us to the low* 10^{-18}, *which is primarily limited by a heating problem in Lu-1.*

We explicitly gave a very crude estimate and pointed out that it made very little difference to the total uncertainty. The full sentence makes clear our position, and it is distinctly not that implied by the reviewers. We now give the total systematic uncertainty as 8×10^{-19} (which is the higher uncertainty of the two clocks, rounded up) exactly as in the quoted paragraph and have added a footnote to Table 2 to clarify explicitly this assumes a BBR temperature of 35(10) °C.

Finally, it would be useful to include the mean values of the comparisons in the last column of the table, for the similar reason we mentioned earlier.

Agreed, updated to include mean and separate statistical and systematic uncertainties for all reported measurements.

On lines 109-113, the authors report 7 nT flicker (1/f) instability of the magnetic field for a 1-100 s range. The loss of contrast is also attributed to this noise. Although the HA-HR protocol largely suppress the quadratic Zeeman effect from the hyperfine splitting, depending on the interrogation scheme, there may still be residual sensitivity to the time-correlated noise such as oscillations at the clock cycle period or 1/f noise. In [T. Lindvall et al., Metrologia 60, 045008 (2023)], it was reported that the servo error can be as large as several parts in 10^{-18} for about 10 pT flicker noise environment with about 100 kHz Zeeman splitting. Here, we have 3 orders of magnitude larger magnetic field noise but similar or larger suppression of field sensitivity (about 20 Hz at 0.1 mT?). How significant could this uncertainty be? It will be common-mode between Lu-1 and Lu-2 but we think it should be included in the overall systematic uncertainty if it is not negligible.

Please see general comments on magnetic field noise. The correct numbers to compare to the Sr^+ system is >4 orders of magnitude reduction in sensitivity and three orders of magnitude increase in noise. Consider 1/f, and cycle-synchronous magnetic noise separately:

1/f : First, as pointed out in general comments, the magnetic noise is not correlated or common-mode between Lu-1 and Lu-2. The servo errors of "several parts in 10^{-18} for 10 pT flicker noise" in T. Lindvall 2023 come from Fig 7 and 8 of that reference and stem from the use of a 'normalized' discriminator [e.g. Dube PRA 92:042119, 2015], which we do not, and have never, used. These bias effects are further exacerbated in a 'six line' servo as used in Sr^+, and by us in Ba^+, over multiple lines of differing sensitivities.

Additional Supplemental Material now details our B field servo implementation, which is a basic two line servo on a single Zeeman pair with 'unnormalized' discriminator, unit gain, and secondary integrator. Previously the mean differential magnetic field was somewhat arbitrarily assigned an uncertainty of 1 nT as it was anyway not a significant contribution. An objective consideration of the average magnetic field uncertainty as constrained by servo projection noise is $\sim 2\times10^{-19}(\tau/\mathrm{s})^{-1/2}$ as given in the original manuscript's methods, and after 200 hours is $\sim 10^{-22}$ which would correspond to an uncertainty in the average field of 10 pT. Hence the differential quadratic Zeeman uncertainty is now given as < 0.01 [E-19] in the table.

Cycle synchronous : This would be technical noise cyclic with the interrogation which is unrelated to background field noise level. The relevant scale is the linear field sensitivities and as noted these are >4 orders of magnitude lower than for alkaline earth ions. To include an effect in the error budget the source must be both plausible and quantifiable. The only plausible source we can think of for synchronous magnetic noise is if magnetic fields are somehow induced by the pulsing of the microwave sources during the Ramsey interrogations. We had in fact been concerned about this possibility in previous precision measurement work on more magnetic field sensitive states. In those works we specifically tested for this by applying much higher than usual power microwave pulses (off resonant), and from different horns, but never observed any measurable synchronous shift in interleaved self-comparisons on the most magnetic field sensitive states. For the low sensitivity $^3D_1$ clock states, we see no justifiable reason to include this.

Regarding Fig. 2, we could not find how the error bars in Fig. 2c is estimated.

Added comment in Fig.2 caption "Error bars represent the statistical uncertainties which are assumed to be projection noise limited instability". There is no significant difference in the projection noise limited instability for Ramsey times 5, 7.5 and 10 due to loss of contrast at longer Ramsey times, hence 5 s was used after the first three runs once this was determined. The data point presented in Fig. 2c is now also expanded in a Table provided in the Supplemental.

On line 218, the authors predict the magnetic field noise would be reduced if they use a smaller magnetic field. Is the magnetic field noise dominated by the current driver of the coils?

If we use a smaller magnetic field B_0, it is the linear sensitivity ($\frac{d\delta\nu_{\mathrm{Qz,k}}}{dB} = 2\alpha_k B_0$) that changes, not the magnetic field noise amplitude. We have performed experiments with only permanent magnets (no coils) before and the magnetic field noise is the same. All indications are it is dominated by the unshielded ambient background and not the current driver noise.

3. First order Doppler shift

The most significant gap in this work is the omission of the first-order Doppler shift (FODS) systematic. This has the potential to be a significant systematic shift in any clock, and especially in trapped-ion clocks; for example, in the manuscript's reference 37, the FODS was measured to be 4.6×10^{-17}, about 460 times larger than the total uncertainty presented here.

The NIST group in ref 37 observed a large frequency difference when probing from opposite directions and hypothesized that it was a FODS due to charging and discharging from photoelectrons generated by uv photons. Proper evaluation of this mechanism requires careful consideration of the underlying physical timescales. Discharging of the accumulated charge has to come with a timescale. If it is fast with respect to the interrogation time, the initial time-dependent shift has to be much larger as it is suppressed by the ratio of interrogation time to discharge-time. If it is slow with respect to the interrogation time it would require multiple clock cycles to reach a stable limit cycle and the total movement from the equilibrium point is substantial and unrealistic, at least for the large shift they reported.

However, even if the uv charging hypothesis is accepted as credible, the physical prerequisites for uv-induced photoelectron charging are fundamentally absent in our Lu^+ system. Photons from the 350 nm repump laser have insufficient energy to overcome the work function of the metal. Even if one was to argue that surface contaminates could lower the work function, there is minimal scattering from the 350 nm laser. The detection channel at 646 nm is practically closed, so any clock cycle would, on average, generate about five photons scattered at 350 nm, and that is only if the atom is found in the ground state. It could only then come from a weak backreflection from the exit port that just so happens to hit the electrodes upon return. For typical 350-nm input power of <100nW, a 10-ms cooling duration, and 4% back reflection, we estimate at most this would generate <1000 photoelectrons based on a 10^{-5} generation probability.

Dynamics of photoelectrons will proceed on a timescale much faster than the trap drive. So, the trap drive can be considered static for all practical purposes. Depending on the rf phase at the time of generation, photoelectrons will either be reabsorbed on the trap electrodes from whence they came or expelled from the trap with little deflection in their trajectory. Those leaving the trap are unlikely to accumulate on insulating surfaces, which, by design, do not have a line of sight to the ion. Hence, even if there is such an accumulation, the field they generate will be attenuated by the amply grounded conductors surrounding the trap.

Our same-species comparison inherently bounds dynamic effects. For a large FODS, of any source, the dynamic motion would have to be identically synchronized with the same magnitude and spatial projection, across two independent traps to evade detection. We consider such a scenario to be highly improbable.

A FODS is induced by any relative motion between the ion and the laser beam path (more specifically, the retroreflecting mirror used as phase reference, see manuscript Fig. 1a) that is synchronized with the clock interrogation cycle. To shift the clock frequency by 1×10^{-19} requires only $10^{-19} \cdot c \approx 3 \times 10^{-11}$ m/s or 0.03 nm/s, less than 10^{-4} of the clock laser wavelength per second.

While the details of the ion trap in ref. 37, which trapped two ions rather than only one, are different from the work presented here, there remain many possible sources of a FODS at the level of 10^{-19}. Some that come to mind include:

- Thermal expansion/contraction in trap or optical materials due to laser or microwave absorption;
- Microwave-induced forces, including electrodynamic forces from eddy currents or ponderomotive or dielectrophoretic forces on trap or vacuum structural materials;
- Lorentz forces due to changing currents in the trap environment e.g. from turning on/off AOMs, changing supply current to MW amplifiers, etc;
- Vibrations induced by shutters or other mechanical elements on the laser table;
- Coulomb forces due to charging of surfaces near the ion, e.g. from thermal or photoelectrons induced by laser beams.

In fairness to the authors, this oversight is not unique to the present work or the Lutetium clock. Several other atomic clock uncertainty evaluations published in the past few years have also neglected discussion of the FODS. However, it is incumbent on the authors to be complete in their uncertainty evaluation, especially when claiming a dramatic improvement in systematic uncertainty over any other atomic clock and especially given the thorough attention to detail in evaluating many other systematic effects.

An order-of-magnitude estimate can illustrate how this is a serious issue which needs to be measured or constrained rather than ignored. Taking thermal expansion of the trap materials as the example, and using the published dimensions of the BeCu trap and endcap rods at ≥2 mm long, we see that thermal expansion by 1.5×10^{-8}/s is sufficient motion to give a 10^{-19} FODS. Using average BeCu CTE of 18×10^{-6}/K, this requires a temperature change of 0.8 mK/s (or 4 mK averaged over a 5 second Ramsey time). Using a published rod diameter of 0.45, BeCu density of 8.3 g/cm3, and BeCu specific heat of 0.4 J/gK, this would require 4 μJ deposited in the rod, or an average of 0.8 μW over the Ramsey time. Energy deposition at this level is not an outlandish possibility; it could come from, for example, backscattered Doppler cooling light, or induced currents from the HA or magnetic field tracking microwave drives.
[Note that published trap dimensions for this estimate are taken from Kaewuam et al., J. Mod. Optics 65, 592-601 (2018). Also note that it took clicking through the references of at least five publications on Lu+ clocks to reach this information; the authors should consider adding a more direct reference with trap details to lines 79-80.]

A fully dimensioned trap figure is in one of the already cited references [45] and described in the text of another cited reference [18]. These are now additionally cited where the trap is introduced.

To be clear, this is a rough estimate, not taking into account how the trap electrodes are attached to their support structure, the relative direction of the trap and clock laser, the thermal conductivity out of the trap rods through bias connections (and through spacers, although Macor is a strong thermal insulator), etc. The point is just that the amount of energy required for 0.03 nm/s of ion motion is very small, and could realistically be delivered as thermal expansion or electromagnetic forces.

To lead to a FODS, this motion must be synchronized with the clock probe cycle, rather than averaged over. Laser-induced effects would be synchronized in this way, since detection, cooling, and state preparation occur in a block between probes; the HA microwaves are more evenly spread across the Ramsey time, but microwave duty cycle differences between the clock probe and auxiliary measurements also induce a level of synchronization. Effects due primarily to auxiliary measurements or intermittent re-cooling after collisions would be synchronized in the same way.

Regarding these additional sources of synchronous motion proposed, we can only note that these are presented as rather vague hypotheticals without a concrete, realistic scenario to formulate a rigorous quantitative evaluation. However, we offer the following points for consideration regarding the viability of the proposed mechanisms:

- We do not consider the suggested mechanism of absorption of microwave or optical power to warm the trap to be feasible. It would certainly need to be differential across the trap structure to result in a first order Doppler shift, which rules out light scattered from the ion, and moreover metals are reasonable reflectors at lutetium wavelengths. That we detect by counting photons collected from a large numerical aperture which are detected with high quantum efficiency gives a sense of scale. Metals are also excellent reflectors at microwave frequencies so that is also not a realistic scenario, more so for the modest powers (mW) needed to drive the transitions in the manner that we do.

- The numbers given by the reviewer to claim the proposed mechanisms are not outlandish are not representative of the experiment. The reviewer derives an energy absorption of 4 μJ in a single interrogation to have a reasonable effect, and then gives 0.8 μW by dividing by the interrogation time. However, all lasers are off during Ramsey interrogation. The lasers are on for a timescale of $\sim$ 10 ms and, in our system, total optical energy into the chamber per interrogation is $\sim 10\,\mu$J primarily from the clock laser. Given that most of this laser power ($\sim 98\%$) exits the chamber and metals are reasonable reflectors at the wavelength of interest, any realistic estimate of the absorbed energy is orders of magnitude below the proposed threshold.

- Lorentz forces between current carrying conductors at incommensurate frequencies e.g. rf vs microwave or rf vs dc will average to zero over the oscillation period. RF currents from the trap drive are stable as required for useable trap operation. Thus they may provide a constant stress but not a dynamical strain. Moreover, a simple estimate of $\Omega_{\mathrm{rf}} C_{\mathrm{trap}} V_{\mathrm{rf}}$ for currents in the electrodes treated as rods clamped at either end of any reasonable dimension would give maximum deflections of $\sim 10^{-14}$ m.

- As noted above and in the supplement analysis, whatever mechanism one proposes for a first order Doppler shift, it has to have a timescale. For the mechanisms proposed, the timescale would be much faster than the interrogation time as lasers and microwave sources are off for the vast majority of the interrogation time. Thus any shift would have to be unrealistically large as it would be suppressed by the ratio of the timescale of the effect to the interrogation time. This is exactly the point of failure with the uv charging hypothesis in ref 37.

We note that probing from opposite directions using counter-propagating clock lasers provides a technical solution to the FODS issue, as demonstrated in ref. 37. Alternatively, self-comparisons with different Ramsey times and auxiliary measurement duty cycles will constrain many possible sources of an FODS, but would be limited by the statistical uncertainty of the self-comparison. (The correlation spectroscopy data presented here already constrains any differential FODS to be $\leq 7 \times 10^{-19}$.) This effect must be accounted for and incorporated into the uncertainty budget before publication.

Any self-comparison would be limited to much shorter interrogation times than the correlation spectroscopy data and therefore cannot give statistically meaningful information to bound FODS. Because we observe no evidence of a FODS at the limit of our precision, the only way to further constrain it would be a significantly longer measurement campaign.

To incorporate a purely hypothetical FODS into the uncertainty budget lacks scientific utility. If such a shift were to become resolvable, then as noted by the reviewer, it can be readily suppressed by the known and simple technique of counter-propagating interrogation.

In the revised manuscript we consider two sources of non-cyclic FODS for which we have plausible mechanisms and which we can empirically bound. However, since a 1×10^{-19} shift requires an average velocity of 3×10^{-11} m/s for the 200 hour measurement duration, this would correspond to over 20 μm of displacement. This is not a small displacement for a heavy static system on an optical table in a climate controlled environment.

Non-synchronous motion of the ion with respect to the trap: In [17] we demonstrated an unprecedented control of micromotion that allowed us to reach the limit imposed by intrinsic micromotion fundamental to ion trap operation. In that work we explicitly showed the variation of micromotion compensation over time, which is exactly the consequence of ion movement and hence a bound on first-order Doppler shifts. This was well below the uncertainty given in our error budget and the reason we did not include it previously. From the micromotion compensation measurements made throughout the campaign, we now estimate the average FODS by linear interpolation of the inferred displacements between EMM measurements. The presumed physical mechanism driving these displacements is drift in the reference voltages setting the potentials.

Non-synchronous differential motion of the apparatus / optical phase: When we measured the differential AOM phase chirp out-of-loop, we also measured the longer term drift of the differential optical phase, which is presumed due to thermal fluctuations of the unstabilized part of the path. This shift was also negligible and therefore was not included in the previous report. We now include this data and the corresponding instability in methods.

4. BBR assessment

The extraordinarily low differential polarizability of the Lu+ clock transition means that the trap temperature evaluation is much less stringent than for any other atomic clock. However, the temperature uncertainty is still relevant and worthy of some attention, given the small total uncertainty. For example, the 1.6 K uncertainty for Lu-2 still constitutes a 4×10^{-20} systematic uncertainty and would be the third largest source of uncertainty overall.

Moreover, it's not entirely clear what the 0.6 (1.6) K uncertainties represent. The lutetium ion doesn't see a single blackbody spectrum, but a combination of blackbody radiation contributions from the trap electrodes, support materials, and vacuum chamber, which emit at a range of different temperatures. With the uncertainty given being equal to the difference between the vacuum chamber and ion trap, it seems likely that it's intended to encompass this full range of blackbody temperatures. This is a valid, well established way to ensure the evaluation is accurate without detailed modeling of the BBR contributions from different surfaces, accounting for their emissivities, etc. However, if the given uncertainties are accounting for this spectral range, then the measurement uncertainty is missing from the evaluation. Alternatively, if the given uncertainties are estimates of the measurement uncertainty, then the spectral range of the blackbody radiation still needs to be accounted for.

The dummy-trap technique used here is in general a valid way to evaluate the blackbody environment. However, there are some additional nuances compared to actually measuring the temperature inside a clock during operation, which introduce additional sources of uncertainty.

1. To make the single measurement of δT_k a valid evaluation of the range of blackbody temperatures seen by the ion, it should be measured at the highest-temperature part of the environment; otherwise, it's not guaranteed that the $\pm 100\%$ uncertainty includes the full range.

1.1. If dielectric loss in the spacer is the largest RF-induced heat source, then this would be the spacers themselves or the endcap rods. However, only the temperature of the end block is measured. Taking the end block as a single temperature should be valid due to the high thermal conductivity of copper, but

Alumina and Macor have thermal conductivities that are 15 and 270 times lower (respectively). (See e.g. the manuscript's reference 14, where the trap support structure was thermally anchored but the electrode temperature changed significantly with RF applied). Have the authors modeled or considered gradients in the spacers?

The gradients in the spacers are modeled in the revised supplemental.

1.2. Additionally, in some ion traps dielectric loss in vacuum feedthroughs can be larger than in spacers, resulting in higher temperatures on the vacuum chamber near the feedthrough. (See discussion in Doležal et al, Metrologia 52 842 (2015).) In others, resistive loss can be higher, whence heat would conduct through to the BeCu trap rods (see e.g. Lindvall et al, PRApplied 24 044082 (2025)). Because the spacers are made from thermally insulating materials, the BeCu trap rods could be significantly hotter than the end block. Were measurements made at these other positions, or is there a way to constrain that the end block does in fact give the highest temperature seen by the ion rather than the trap rods or the RF vacuum feedthrough flange?

The resonator Q drops by only 7% when attached to the feed-though only, compared to a factor 2-4 reduction when the trap is attached. For the simulations in the revised manuscript, we assume 100% of the rf power is dissipated in the spacers to determine the worst case rise in electrode temperature relative the end block.

2. There is no uncertainty contribution for how well the dummy trap is expected to replicate the real trap. Going based on the drawings in Fig. A3b without additional information, it looks like the physical trap assembly should be very repeatable and not introduce additional uncertainty. However, other possible differences between the dummy and real traps should be considered.

2.1. The loaded Q for the dummy trap resonator is 25% higher for Macor and 40% higher for alumina than the real traps. This would seem to imply more loss in the RF system for the real trap, and therefore a potentially higher temperature. Can the extra dissipation be accounted for? If not, this should be included as a source of uncertainty.

2.2. Does the dummy trap accurately reproduce the RF environment of the real traps, including vacuum feedthroughs and nearby DC/RF electrode bias lines? These could change the distribution of current and therefore temperature in and around the dummy trap.

The dummy trap is Lu-1 from ref [16] which was rebuilt and put back into an identical vacuum chamber. As stated it is identical to the two traps in use, including the entire in-vacuum assembly as it is attached to the rf feedthrough on the vacuum side. DC lines connect at the sub-board several cm from the trap, opposite the rf input side and cannot be expected to influence RF loading.

2.3. The authors observed a large difference in heating between alumina spacers from different manufacturers. What constrains differences between the alumina spacer used in the dummy trap and the real trap are they known to be machined from the alumina with the same impurity composition? The authors consider the possibility that the original spacers may not have been made from alumina as the manufacturer claimed. If there's a concern about manufacturer malfeasance, is it reliable that these spacers were made from the same batch of material? If additional pieces from the suspect manufacturer are available, this could be tested experimentally. If that is not possible, the assumption that the dummy and real spacers share the same composition should be explicitly stated in the text.

It is now stated in the Supplemental explicitly that dummy trap test spacers are extras left over from the original batch orders used to construct Lu-1 and Lu-2. We are certain of this fact, and it is only reasonable to assume that they come from the same material stock, given they were delivered together from a single supplier. We have now had the spacers SEM/EDX tested and confirmed the sample is Alumina, so that speculation has been removed from the manuscript.

3. The authors measure a temperature gradient across the enclosure. What about temperature gradients across the vacuum chambers? For example, a steel vacuum chamber could have a gradient if the stabilized air temperature is different from the laser table temperature (for example due to thermal conduction across or through the bulk of the table, if the stabilized temperature is different from the surrounding room temperature). If such a gradient does exist, it may be appropriate to add it to the RF heating linearly rather than in quadrature, as it would extend rather than shift the range of blackbody radiation seen by the ion.

For both chambers we measured a gradient of 0.44 degrees from top to bottom of the vacuum chamber. We now take half this value as the uncertainty in the offset to the ambient temperature. In the revised treatment we take the lower temperature bound as the ambient minus 1σ and the upper bound as the ambient plus 1σ plus rf heating, in effect extending the range as suggested.

4. The calibration uncertainties of the PT100 sensors do not appear to be specifically accounted for. These should be added in quadrature with each other and linearly with the temperature range discussed above (since the temperature range isn't the width of a Gaussian distribution of possible values, but rather a spread of temperatures which all must be accounted for, it can't be treated as uncorrelated with the measurement

uncertainty).

All sensors were calibrated against a single PT100 sensor so any calibration is fully correlated. The specification sheet gives ±150 mK (at 30 C) but is not explicitly stated in data-sheet if this is $k = 2$ so we assume $k = 1$ as worst case. Calibration uncertainty is fully correlated for all the measurements, both for trap heating and the ambient environment.

The authors should account for these additional sources of uncertainty on the measurement of the temperature, to build more confidence that this evaluation is accurate despite not measuring the actual trap used in the clocks.

4. Magnetic field servo

It seems like a servo error on the magnetic field servo would translate directly into a clock frequency shift. In the data shown in Fig. A1, the out-of-loop measurements constrain the average field value at the 1 nT level, which is a negligible second-order Zeeman shift. However, this only shows one hour of data and servo error can change depending on the behavior of the system being servoed (e.g. drift rate) were similar out-of-loop measurements taken during the comparison, or can servo error be constrained in some other way?

Additionally, we find no mention of the mean value of the difference between the in-loop and out-of-loop magnetic field measurements. Presumably, this implies the difference is smaller than the statistical uncertainty, but this should be stated explicitly.

Additionally, could you explain how in-loop measurement of the magnetic field averages as $1/\tau$ not $\sqrt{1/\tau}$?

The $1/\tau$ averaging is a somewhat subtle consequence of the magnetic field servo implementation which is now explained in the Supplemental. The 'out-of-loop' measurements were not taken during the actual comparison; this was just a separate test to demonstrate the typical magnetic field noise and servo performance, and verify the observed servo behavior matched simulations and our understanding.

A supplemental material now explains details of the magnetic field servo. It includes the fields measurement during the comparison, as well as the applied field corrections by the shim coils, for the longest single run. For the measured Zeeman splitting during this actual run, the measurements continue to average $1/\tau$ for the full duration. We take the out-of-loop measurement as the instability of the field, which is above the servo's projection noise due to the measurement deadtime in the field servo and has been fully modeled with details given in the Supplemental.

5. Second-order Doppler shift A couple of minor questions here. Do the authors have any thoughts as to why the measured temperature is so far different from the calculated Doppler limit, given that the heating rate is quite low? While measurement should be trusted over theory for the trap temperature, an explanation for the discrepancy could help build confidence in the result.

Achieving Doppler-limited cooling on Lu^+ with a ground state consisting of 3 three hyperfine manifolds and 45 substates is non-trivial, and not even easily simulated. We select a laser detuning and intensity configuration which simultaneously gives us reasonable cooling and also gives high saturation of the excited state for efficient state detection.

Even at low heating rates, the final ion temperature and uncertainty is dominated by the heating over the 5 s Ramsey time, so further improving the initial cooling provides little gain for this experiment. We note other cooling schemes are available to achieve Doppler, or sub-Doppler, cooling such as the degenerate Raman scheme used for ground state cooling, or a counter-propagating traveling wave configuration (not yet published).

Additionally, were temperatures and heating rates measured on both radial secular modes, and did they agree?

See response to reviewer #3.

6. Coupling errors

As I understand this section, the coupling error is obtained from a set of three measurements, with one each taken at the beginning, middle, and end of the measurement campaign; and the uncertainty on the coupling systematic is the maximum spread of these three values. Can the measurements distinguish between long-term drift and shot-to-shot or day-to-day variation? Do the authors expect that three measurements will adequately capture the full range of couplings across the measurement campaign? Additionally, is the uncertainty on each individual measurement negligible compared to the spread between measurements, and if not has it been incorporated into the stated uncertainty on the shift? Since this is the second largest source of uncertainty for Lu-2 (and since it also determines the AC Stark shift, see below), a bit more detail would be useful.

While running simulations to generate an additional figure for the microwave and optical coupling errors, we realized a mistake in the previously reported assessment of the optical coupling error. The 'optical coupling error', equation (12) in the original manuscript, applies only for HA without Hyper Ramsey. Inclusion of Hyper Ramsey further suppresses this shift, similarly to the AC stark shift, as is in fact already pointed out in the supplementary material of ref [16] where expressions are given for its evaluation. For this experiment (which includes Hyper Ramsey) the 'optical coupling' contribution is negligibly small ($\ll 10^{-24}$) and so has been removed from the uncertainty budget. We note this makes no difference to the final combined uncertainties.

The variation in measured 'optical coupling' is still used to bound the range of laser intensity variation as it applies to the ac Stark shift.

Both microwave and optical coupling rates are measured by fitting the respective Rabi flopping curve with typical $\leq 0.2\%$ measurement accuracy. This is now noted in the methods. Clearly three measurements alone cannot reasonably distinguish variation within the campaign and provide only minimal information on the drift, but we can have reasonable confidence from the typical behavior when operating the apparatus at other times. During intense experimental setup phases, the microwave and optical Rabi rates may be measured many times throughout the day. The laser intensity at 848 nm is actively stabilized and typical coupling variation is within $< 1\%$ range over the day and day-to-day. Microwave coupling rates are typically exceptionally stable, in $< 0.5\%$ range, as long as the microwave horn and all metal pieces near the chamber are stationary. The 1.5% microwave coupling and 2.5% microwave coupling uncertainty given are due to a larger than usual deviation of a single measurement in Lu-2. We expect using the full range as the standard deviation to still be a conservative estimate to capture the range of values throughout the campaign.

Even though 'microwave coupling' is the largest uncertainty in the differential uncertainty budget, we have taken a conservative bound on it by using Eq. 11, which is for the worst case scenario in microwave detunings. We could provide a more rigorous evaluation incorporating the known microwave detunings Δ_k from the experimental data; this will inevitably reduce the 'microwave coupling' uncertainty and the total comparison systematic uncertainty. Given the total systematic uncertainty is already well below the comparison measurement precision, we retain the conservative estimate.

8 AC Stark shift

The uncertainty here appears to be based on the "observed variation in the optical couplings", i.e. the same set of measurements discussed in the "Coupling Errors" section. If so, it may not be appropriate to add these two effects in quadrature in the uncertainty budget. The "Optical coupling" entry for Lu- 1 is negligible but for Lu-2 it may be more appropriate to linearly add the uncertainty contributions due to optical coupling errors and uncertainty on Δ_S , and then add in quadrature the contribution from uncertainty on the measured heating rates.

As explained in the comment above, 'optical coupling' should not be in the uncertainty budget and has been removed. Even so, in the previous assessment the "Optical coupling" entry for Lu-2 was small (4×10^{-21}) compared to other shifts such that adding linearly to the 'ac-stark' shift would make no difference to total uncertainty (9.8×10^{-20}) to the significant digits given. The reviewer may be mistakenly referring to the Lu-2 'microwave coupling' shift (5.2×10^{-20}), but this is expected to be uncorrelated with other shifts.

9 Differential AOM chirp

"Differential AOM chirp" section seems to discuss about two different measurements. The first part is on a proxy of the out-of-loop measurement on the clock laser phase, which is closely related to FODS we concerned earlier. However, we could not find the result of this measurement. Although this measurement does not include the sources of FODS on the trap, it would give some idea. What is the result of this measurement?

We have now included this data and expanded the discussion.

Response Reviewer #3

This manuscript demonstrates a landmark achievement in optical timekeeping. The authors set a new benchmark in systematic uncertainty for an optical clock, reaching 1.1×10^{-19} for their $^{176}Lu^+$ system. The central and most significant result is the direct validation of this accuracy via a frequency comparison—enabled by correlation spectroscopy—between two independent clocks, showing agreement at the level of $-2.4 \pm 5.8 \times 10^{-19}$. This represents the first demonstration of a clock's performance below the 10^{-18} level validated by a same-species comparison.

The result is a major advance for the field. While the use of 176Lu+ and techniques such as hyperfine averaging and correlation spectroscopy builds on prior work by this group and others, the achievement of validated 10^{-19}-level performance is unprecedented, representing a significant technical milestone. Its importance lies in providing the first conclusive, self-consistent validation at this level, thereby meeting the stringent 2×10^{-18} threshold set by the CCTF for redefining the SI second.

That said, the strength of this work lies in its profound technical depth and its role as a definitive benchmark for the atomic physics and metrology community. While the result is monumental, its core impact is vertical—deeply advancing a specific field—rather than horizontally broad, which journals such as Nature often emphasize. I believe this work certainly deserves publication as the next step beyond recent PRL papers by Aeppli et al. (Sr, 8.1×10^{-19}, PRL 2024), Marshall et al. (Al^+, 5.5×10^{-19}, PRL 2025), and Zhang et al. (Ca^+, 4.4×10^{-19}, PRL 2026), and would be most appropriately published in a specialized, high-impact physics journal such as Physical Review Letters.

We thank the reviewer for their concise and accurate review, acknowledging the merit of both of their two main points of criticism. However, we disagree on the characterization of this work as a small step beyond those three papers indicated, none of which include the self-consistent validation component, and which distinguishes this work. The revised manuscript highlights the critical importance of this in the context of the current state of the precision metrology field.

The experimental approach is sound and the data quality is very high. The presentation of the systematic uncertainty budget is clear and detailed. However, two aspects of the methodology require strengthening:

(1) Radial Heating Rate Anisotropy: Figure 6A reports the heating rate for only one of the two radial motional modes. In linear Paul traps, the two radial directions are not necessarily equivalent due to potential asymmetries in electrode geometry or the surrounding electrical environment. Heating rates can differ significantly between these modes because of variations in coupling to electric field noise, patch potentials, or dielectric losses. *Request:* Please provide the measured heating rate for the second, orthogonal radial mode. Both radial heating rates should be included, for example in the supplementary information.

The second radial heating rate was not measured because ground state Raman cooling is needed, and the Raman beams only coupled to one of the radial modes. There are technical challenges with reconfiguring to cool to the other mode, though it is possible. While we agree this should in principle be included for completeness, it would be time consuming to go back and measure this now, and disruptive to other ongoing experiments with the apparatus. Thus we have to balance the effort required for this against realistic considerations of how much different the radial mode heating could be and potential influence on the final uncertainty.

Radial heating rates are expected to scale with secular frequency due to the electronic noise spectrum as $\dot{T} \propto \omega^{-\alpha}$ where varying ('anomalous') scaling coefficients in the approximate range $\alpha = [0.5, 1.6]$ (Rev. Mod. Phys. 87, 1419 (2015)) have been observed. Given our radial mode frequencies differ by only 7% for Lu-1 and 9% for Lu-2, the frequency scaling effect in the radial heating rates ($\dot{T}$) is expected to be smaller than or comparable to the reported measurement precision on the radial heating rates.

Heating rates have been extensively studied but where heating anisotropies have been reported have been in surface traps for which ions are much closer to the surface, often have nearby dielectric surfaces, and anisotropic geometries. It is also well known that these 'anomalous' surface effects fall off rapidly with distance from the electrode. Consider our macroscopic trap: the ion is $\sim$600 μm from the rods, the trap has perfect radial xy-symmetry and there is no line of sight to any dielectric in the radial plane. To have a difference in radial heating rates, beyond the frequency scaling discussed, could only come from a noise source on an electrode which somehow couples to one mode despite subtending a significant solid angle. We consider the possibility highly implausible.

We are not currently aware of any reports in the literature of substantial anisotropic heating (beyond what can be attributed to frequency scaling) in a similarly macroscopic and symmetric linear Paul trap, making this scenario highly improbable. However, we have updated the methods section to explicitly note that the second radial mode has not been measured, and our assumptions and considerations made for the unmeasured mode. For the unmeasured mode we now double the uncertainty in the heating rate for that mode; although this has no influence on the final SODS result.

(2) Blackbody Radiation (BBR) Shift Assessment: The BBR shift is a dominant systematic uncertainty in high-accuracy optical clocks. Currently, the Methods describe its evaluation based solely on temperature measurements of the two "end blocks." This approach is likely insufficient and may underestimate the true uncertainty, because the temperature at the end blocks need not represent the temperature of the electrodes at the ion's position—the relevant parameter for the BBR shift. In both the Lu-1 and Lu-2 traps, the RF and DC electrodes experience ohmic heating from applied voltages and likely follow different thermal pathways, meaning the electrode temperature could be substantially higher than that of the more massive end blocks.

Request: The authors should provide a more rigorous, quantitative assessment of the BBR shift. This should include:

(i) Enhanced thermal analysis, which could be achieved via direct temperature measurements closer to the trap electrodes, or through a well-validated finite-element thermal model that incorporates RF/DC ohmic heating and the detailed trap geometry. Alternatively, a holistic evaluation might involve measuring the temperature and emissivity of each material inside the vacuum chamber, along with the solid angle subtended at the ion.

(ii) An updated uncertainty estimate for the BBR shift based on a realistic upper-bound for the electrode temperature, rather than relying solely on the end-block temperature. An explicit discussion of the likely thermal gradient between electrodes and end blocks is necessary, and its contribution should be included quantitatively in the overall uncertainty analysis.

This is a valid point. We have now estimated a realistic upper bound temperature for the electrodes using an approach along the lines of the reviewer's suggestions. This can be found in the new Supplemental Material.

Minor Comments: (1) The "Hyperfine-Averaged Hyper-Ramsey" sequence used to suppress probe-induced shifts is clever. To assist other groups in implementing similar protocols, could the supplementary material provide further details on the exact pulse sequence—for example, a diagram or a list of phases—along with the derived robustness conditions against common-mode pulse-area errors?

The exact pulse sequence with phases is already in the main diagram. The timing parameters used here are given in methods. The supplemental of [16] extensively derives general analytic expressions for the coupling (pulse area) errors in both the microwave and optical pulses for HA both with and without HR. The expression given in the methods for the microwave coupling errors comes for this supplement. Since we have nothing further to add to that analysis, we refer the reader to the supplement of [16].

The ac-stark methods section now includes a discussion about averaging HR and time-reversed HR sequences to effectively suppress thermal effects in HR. Even though we don't use this method here since the ac-stark shift is already sufficiently small, it may be of interest to the reader who may wonder about the differences in ordering of the HR pulse sequence here as compared to previous work, as another reviewer inquired.

(2) The long-term stability of the (α_{QZ}) coefficient over three years (Fig. 3) is remarkable and is a key pillar of the low uncertainty.

We thank the reviewer for noting this. Although the total magnitude of the QZ shift is large relative to other shifts in Table 1, the QZ coefficient is quite small compared to other clock systems and not demanding on the magnetic field requirements to achieve low uncertainty.

Were these measurements taken under identical trap conditions (e.g., ion position, RF power)?

The first measurement was made by comparing the two systems in ref[16] at different magnetic fields. After that, the labs were renovated, many systems upgraded, and the Lu-1 trap rebuilt. The last two measurements were taken in the existing systems, but one year apart and with different RF-drive frequencies/resonators and secular frequencies.

Could the authors comment on any observed sensitivity of (α_{QZ}) to trap alignment or other parameters that future experiments should control?

α_{QZ} itself does not have any sensitivity as it is an atomic property and the QZ shift depends only on the dc (and co-linear ac) magnetic field. The measurement of α_{QZ}, however, is done by comparison of independent systems so all other systematics must be accounted for in both systems at their respective operating fields. By going to a high field (a few mT) the QZ shift is much larger than other systematics such that high precision determination of α_{QZ} is possible with lower demands on the measurement precision and systematic uncertainties. The consistency of the measurements over three years does give us high confidence in the evaluated uncertainties and effectiveness of HA over a large magnetic field range.

(3) The BBR shift is the main area where the uncertainty analysis could be made more rigorous. Additionally, could the authors clarify whether any potential correlations in environmental noise (e.g., magnetic field fluctuations) between the two independent traps were considered in the uncertainty of the frequency comparison?

See general comments on magnetic field noise, all evidence indicates magnetic field noise is uncorrelated in the two systems. For other environment factors, e.g. micro-motion, stray electric-field gradients (quadrupole shift), rf ac-Zeeman shift etc. there is no reason to expect correlation in different traps and certainly it is not assumed: the corresponding environment perturbations are measurably quantified for both systems independently. Temperature fluctuations of the ambient environment are correlated, but given the low BBR sensitivity, are far below the measurement precision.

(4) Please update the references to include formal journal citations for entries currently cited as arXiv preprints that have since been published.

References updated.

Response Reviewer #4

The authors report a frequency comparison between two independent $^{176}Lu^+$ single-ion clocks using correlation spectroscopy with a shared clock laser. This technique enables an instability of $4.8 \times 10^{-16}(\tau/s)^{-1/2}$. The systematic uncertainty is evaluated at 1×10^{-19} based on measured field sensitivities and coefficients, while the total measurement uncertainty of 5.3×10^{-19} is currently limited by statistical instability.

We thank the reviewer for taking the time to review this manuscript. The comments on the language and structure of the manuscript are well received and the main text of the manuscript has been substantially revised for accessibility to broad readership.

1) The reported relative frequency agreement is dominated by a statistical uncertainty of 5.3×10^{-19}, which exceeds the quoted systematic uncertainty. While the result is consistent with agreement at the 10^{-19} level, the measurement does not statistically resolve agreement at that level. Given the broad readership of Nature, the authors should clearly clarify this distinction and

ensure that claims of 10^{-19} performance are framed as inferred or consistency-based rather than experimentally demonstrated.

The language and framing has been revised to be very clear on these points.

2) "Here we report two 176Lu+ single-ion optical clocks, each with fractional frequency uncertainty near 1×10^{-19}, a five-fold improvement over the highest accuracy reported to date." The use of correlated spectroscopy with a shared interrogation laser removes laser noise by design, which is appropriate for probing atomic agreement and common-mode rejection, but does not capture the performance of two independent clocks, for which the clock laser is an integral component. The authors should clarify that the reported agreement reflects atomic reference consistency under shared-laser interrogation, rather than standalone clock performance.

We respectfully disagree. In an ion trap system, the only influence the laser will have on clock performance is limiting the possible interrogation time and hence stability. Correlation spectroscopy explicitly removes the laser from consideration and demonstrates the degree to which two independent atomic references agree, and hence the accuracy that can be achieved using this reference.

3) This reviewer finds the following text misleading: "We demonstrate interrogation times up to ten seconds and a comparison instability of $4.8 \times 10^{-16}(\tau/\mathrm{s})^{-1/2}$, nearly a factor of 3 improvement over previous work [16]. Comparing optical references via correlation spectroscopy has $\sqrt{2}$ higher instability than comparing independent clocks, and a factor of two higher instability than an individual clock, for the same interrogation time. It follows that these $^{176}\mathrm{Lu}^+$ systems are currently capable of $\sim 2.4 \times 10^{-16}(\tau/\mathrm{s})^{-1/2}$ clock instability if combined with state-of-the-art laser stabilization technology [28]. This is an improvement over the lowest instability demonstrated in an ionic frequency standard [12], and would enable instability below 10^{-18} in less than one day."

The final inference is problematic. The projected single-clock instability relies on assumptions that are not met experimentally: namely, that a single clock can be interrogated for the same duration as enabled by correlation spectroscopy (which is unlikely given that trap lifetimes of typically of the order of 1 s due to micromotion and magnetic field noise effects), and that state-of-the-art cryogenic laser stabilization (absent from the present system) is available. As written, this risks conflating demonstrated comparison performance with hypothetical single-clock capability. As a result, this statement should be revised or removed.

We are aware of no work where micromotion limits the clock coherence time. Consider the polarizability measurements in Sr^+ [PRL, 112(17):173002, 2014.] and Ca^+ [PRA, 99(1):01140, 2019] where very high modulation depth micromotion is intentionally induced without degraded stability. In the work here, EMM shifts are far smaller and stable. Furthermore, EMM in the independent chambers is in no way correlated, so if it was a limit to the optical interrogation time, it would also limit our correlation spectroscopy time.

For the magnetic field, as discussed in the general comments and revised manuscript, all evidence indicates that dephasing due to the magnetic field noise is uncorrelated in the two separate systems. The reviewer's statement '*which is unlikely given that trap lifetimes of typically of the order of* $\sim$*1 s due to micromotion and magnetic field noise effects*' is in contradiction to our observation of > 10 s coherence times on the microwave clock transitions for systems individually.

Correlation spectroscopy only rejects common mode phase fluctuation from the laser and common mode environmental influences. Insofar as the environmental influence limiting the coherence time (magnetic field) is not common mode, it follows that this stated clock stability would be achievable if the optical coherence were available. We stand by that statement but have moderated it by removing any claims of 'best' stability, as it is indeed hypothetical contingent on state-of-the-art cryogenic laser stabilization, and added the qualification 'Insofar as magnetic field noise is uncorrelated'. It is however an important point to make that ultimately ion clock instability can be limited either by the clock laser or by the atomic coherence through, for example, the excited state lifetime or environment perturbations. That we are able to demonstrate this level of atomic coherence between ions in two separate traps in a relatively noisy magnetic environment is demonstrable evidence of the single-clock capability that can be achieved in Lu^+.

While the authors claim and improvement of a factor of 3 as compared to previous demonstrations, the comparison is relative to the author's previous measurements [16], and does include a comparison to the state-of-the-art from other groups. This self-comparison needs to be clarified in the text, or the claims should be made against the previous work of the broader community:

- Aeppli, Alexander, et al. "Atomic clock frequency ratios with fractional uncertainty $\leq 3.2 \times 10^{-18}$." arXiv preprint arXiv:2512.21428 (2025).
- Kim, May E., et al. "Improved interspecies optical clock comparisons through differential spectroscopy." Nature Physics 19.1 (2023): 25-29.
- Clements, Ethan R., et al. "Lifetime-limited interrogation of two independent Al+ 27 clocks using correlation spectroscopy." Physical review letters 125.24 (2020): 243602.

All three works are now cited in the context of the state-of-art for comparison instabilities between independent systems, but it must be noted that the final two references are not actually comparisons in that no frequency ratios are reported. The frequency ratios reported in Aeppli et al [BACON25 published after our submission] and the earlier BACON21 now feature in the revised manuscript.

4) The language used in much of the text, but especially in the Experiment Set up and Results sections, seems to be targeting an audience very familiar with atomic clock operation. The authors rely heavily on undefined technical terms such as "hyper Ramsey spectroscopy", "correlation spectroscopy", "Lamb-Dicke" parameters, "hyperfine averaging" etc. As such, in its current form is not written in such a way as to be accessible to the broad readership of Nature. At the least, both sections should be rewritten, replacing jargon with accurate but simpler language explaining clock operation, notable differences or similarity to other groups and systems, and a clear explanation of how the Lu+ clock operation enables advantages in performance as compared to previous reported measurements.

- For example, correlation spectroscopy is used to probe both ion samples simultaneously with a shared laser, which helps to suppress laser noise and enables longer interrogation times (Ramsey times).
- Hyperfine averaging: averaging over magnetically sensitive transitions helps to suppresses the first-order Zeeman shifts on the clock transition frequency.
- These types of explanations would help orient readers prior to Eq. (1).

We thank the reviewer for their suggestions and the manuscript has been updated accordingly. Technical details are now moved entirely to methods and Supplemental material to allow for greater discussion.

We note that hyperfine averaging suppresses second order Zeeman shifts of $^3D_1$; there are no first order Zeeman shifts on any transitions used for clock operation.

The work is technically strong and addresses an important question in optical clock performance. However, for Nature's broad readership, the current framing overstates the level of experimental validation and relies heavily on specialist language. With clearer distinction between inferred systematic control and statistically resolved performance, explicit discussion of the implications of correlated spectroscopy (a method developed elsewhere), and more accessible description of the experimental protocol, it is unclear whether the gains in performance represent sufficient novelty for publication in Nature. As written, and given the current results in comparison with the state-of-the-art, this manuscript seems more suitable for PRX or Nature Physics.

In the revised manuscript, we now provide a rigorous definition of 'validation' and clarify the vital role of unit-ratio comparisons in substantiating evaluated uncertainty budgets. Given the impending redefinition of the SI second, establishing evidence-based performance for optical frequency standards is a critical milestone. We believe that highlighting these methodological distinctions provides essential context that will be of significant interest to the wider scientific community.

I. SELECT NUMBERED REFERENCES FROM FIRST SUBMISSION

[2] Dimarcq, N. et al. Roadmap towards the redefinition of the second. Metrologia 61, 012001 (2024)

[6] Sanner, C. et al. Optical clock comparison for Lorentz symmetry testing. Nature 567, 204–208 (2019)

[7] Takamoto, M. et al. Test of general relativity by a pair of transportable optical lattice clocks. Nature Photonics 14, 411–415 (2020)

[15] McGrew, W. et al. Atomic clock performance enabling geodesy below the centimetre level. Nature 564, 87–90 (2018).

[16] Zhiqiang, Z., Arnold, K. J., Kaewuam, R. & Barrett, M. D. 176Lu+ clock comparison at the 10^{-18} level via correlation spectroscopy. Science Advances 9, eadg1971 (2023)

[17] Arnold, K. et al. Enhanced micromotion compensation using a phase-modulated light field. Physical Review A 110, 033115 (2024)

[18] Lee, M. D. K. et al. Precision measurement of the $^{176}Lu^+$ 3D1 microwave clock transitions. Physical Review A 113, 012805 (2026)

[19] Barrett, M. & Arnold, K. Analysis of collision shift assessments in ion-based clocks. arXiv preprint arXiv:2515.05474 (2025)

[37] Brewer, S. M. et al. "^{2}7Al$^+$ quantum-logic clock with a systematic uncertainty below 10^{-18}. Physical Review Letters 123, 033201 (2019).

RESPONSE TO REVIEWERS AND EDITORS

We thank the referees for their constructive reports. Following the editor's guidance, all discussion of community practices has been removed from the manuscript, and this response is largely confined to changes made to the text. All changes are highlighted in the marked-up manuscript; line numbers refer to the clean version.

POINT-BY-POINT RESPONSES

Referees' comments (blue), Authors response (normal text)

Response to Reviewers #1 and #2

This review attempts to walk a somewhat fine line. On the one hand, we continue to believe that the experimental work reported here is both high-quality and of significant and broad enough interest to merit publication in Nature. This work marks several significant steps forward for atomic clocks, at a time when the redefinition of the second amplifies the importance of atomic clock progress. We discuss this further below.

On the other hand, this revision has added a substantial focus on the authors' opinions about the state of atomic clock research and the shortcomings of the SI redefinition process. We have significant issues with these opinions and their inclusion in this report. The recommended changes conflate the clock evaluation with the more rigorous process of validation; they would likely not resolve the issues with comparison repeatability cited by the authors; the underlying thesis is applied in a logically inconsistent way; and the proposed criteria are not fully met by the experimental work presented here. We will also discuss each of these points in more detail below.

In sum, we do not believe that the manuscript should be published in its present form, but we hope the authors will revise and continue to seek publication in Nature. We believe that a version of this article which emphasizes the impact of the experimental work will satisfy other referees' concerns about broad interest, without centering the authors' opinions at the expense of their remarkable experimental achievement.

Other points are included at the end of the review, including concerns about the unmeasured secular-motion heating rate, where we note that published works and preprints have reported up to a factor of 5 anisotropy in macroscopic, symmetric traps.

Regarding impact and interest

Both of the other referees expressed the opinion that this result is not of sufficiently broad interest to serve Nature's readership. We disagree for several reasons, which we will summarize here.

First, there is the matter of multiple-record-setting performance. This is the most accurate clock ever constructed, by a factor of more than four. In addition to that evaluation, this work also reports the most precise clock comparison to date (and one of the most precise measurements of anything ever). We believe these records alone carry a certain amount of interest to physical scientists, and papers reporting record advancements in clock performance have been reported in Nature - e.g. Bloom et al., Nature 506, pages 71–75 (2014) reporting the JILA strontium lattice clock.

Second, this is the first comprehensive accuracy evaluation of a lutetium clock. This is an important milestone, since lutetium promises significant advances in accuracy compared to better-established clocks. In this sense, we find it analogous to initial evaluations of other types of clock previously published in Nature, e.g. King et al., Nature 611, pages 43–47 (2022) reporting the first comprehensive evaluation of a highly-charged-ion clock, as well as the many Nature reports of incremental progress towards a thorium nuclear clock. We note that this is especially important at a time when the CCTF is establishing the short-list of atomic clock candidates for the upcoming redefinition of the second; this work makes a strong argument for lutetium's relevance to that process. Perhaps this could be emphasized in a future revision.

Third, the combination of a systematic uncertainty evaluation with a same-species comparison is a significant advancement beyond the single-clock evaluations that the other referees compare to this work. While we disagree (see below) with the authors that same-species comparisons should be required before any clock evaluation can be published, it is indisputable that a same-species comparison makes a clock evaluation stronger; it bounds many types of possible experimental error, should make interspecies comparisons involving the clocks more reliable, and demonstrates clock performance in a more realistic scenario than an evaluation alone. We note that the only analogous work of which we are aware was also published in Nature, as McGrew et al, Nature 564, pages 87–90 (2018) reporting on the NIST ytterbium lattice clocks.

We also hope that the reorganization of the paper, with many technical details moved to the Methods section or supplemental material, and with important references and explanations added to many important points, will make it more accessible to non-specialists.

Issues with the revised manuscript

Having made the case that the scientific merits of this work rise to the high level expected from a Nature publication, we move to reasons we do not believe it should be published in its current form. We continue to have issues with the authors' discussion of "validation", as well as their new criteria for "acceptable" atomic clock evaluations. We do not find the authors have adequately addressed our concerns on this front; instead they have devoted significant new emphasis to their interpretation of "validation", which we find flawed, as well as to their criticism of atomic clock research more broadly.

Before getting into the specifics of this debate, we note that a peer-reviewed research report is not the appropriate forum for such opinion. This would conflate peer review of the research with peer review of the opinion. Large parts of the new "discussion" section are essentially unrelated to the research work being reported here, instead being part of the authors' criticism of the state of the field. Rather than inside a high-profile research report, we believe the community and the authors would be better served if these criticisms were published separately in a journal like Metrologia. That would be a more fitting venue for public debate; and the criteria could be presented as an aspirational standard for clock groups including the authors, rather than undermined by being published alongside a comparison which does not fully meet them.

Commentary on the roadmap has been removed and the roadmap is now only minimally introduced for context. The comparison discussion has been reframed around geodesy, the essential consideration for clocks at this level of precision.

The three criteria are now presented solely as the standard to which we hold our own system, with citation (Arnold, Kyle J., et al."Validating a lutetium frequency reference." Journal of Physics: Conference Series. Vol. 2889. No. 1. IOP Publishing, 2024.) to where the criteria first appeared. The phrasing has been tightened to hopefully avoid misinterpretation while conveying the originally intended meaning.

There are a few separate issues here which we will break down.

Regarding "validation" as a term of art

The authors' new definition of "validation" is significantly less rigorous than the community consensus included in the SI redefinition roadmap, mandatory criterion I.2 (N Dimarcq et al., Metrologia 61 012001 (2024), ref. 2 of the revised manuscript). The core element lacking from the authors' definition is reproduction by independent experimental groups. "Validating" an atomic clock requires reproduction by different institutes, which is an essential check against experimenter error, subconscious bias, challenges from long-distance time transfer or clock transportation and unidentified shifts. This is fundamentally not something that can or should be done by a single institute.

For example, the authors cite the NIST Yb optical lattice clock (ref. 15 of the revised manuscript) as the only other atomic clock which meets their definition of a "validated" accuracy budget. We point them to the appendices of arXiv:2512.21428 (ref. 8 of the revised manuscript). Here, this same group discusses how, because of errors in atomic structure calculations, the systematic shifts applied in ref. 15 were too small by 3E-18 – more than double the total systematic uncertainty evaluated in ref. 15. Clearly, the presence of a same-species comparison with high precision did not guarantee that the accuracy budget was correct.

The error came about from using a calculation in determining the uncertainty as opposed to a measurement. More importantly, the resulting error of 3×10^{-18} applies only to the reports in which it was used, namely (McGrew Nature 564, 87–90 (2018)) and the first BACON result, for which it is only half the reported uncertainty of 6×10^{-18}. So the error made no difference to the reported value. To quote from the paper arXiv:2512.21428: *For comparison, the BACON21 evaluation would yield a lattice light shift of* $-3.2(1.1) \times 10^{-18}$ *for these same conditions, consistent below the* 1σ *level.* It is misrepresentative to apply the 3×10^{-18} to the second reported uncertainty where the quantity of interest was measured and confirmed the result of another group.

Additionally, the degree of independence of the two clocks matters when determining whether the evaluation can be considered "validated". As an example of the kind of unidentified shifts which are not eliminated in this comparison, we note that spectral impurities in a clock laser can cause – and have caused, in optical lattice clocks at least – significant systematic shifts, which would cancel in common mode in a comparison like the one performed here. Requiring that comparisons be between different institutions establishes an objective baseline for what can be considered "independent", which otherwise is highly subjective (as one can clearly see from our review and the authors' response).

This contention over "validation" is in part a semantic issue, since – as we discussed in our previous report – "validation" already has a specific definition in the context of next-generation clock uncertainty estimates. The authors have a legitimate claim that their uncertainty evaluation is better supported by evidence than others which do not include self-comparisons. A different, less loaded term would avoid this conflation of the two definitions – perhaps "confirmation", "verification", or "corroboration". However, both because of the common institution and because of the reduced level of independence of the two clocks, we continue to hold that this comparison doesn't reach the level of a fully "validated" frequency reference.

We have adopted the referee's suggested terminology. "Validation" is replaced by "verification" throughout, including the title. The use of "validation" (Discussion, lines [241–247]) refers to the roadmap's inter-institute definition, explicitly distinguished from the within-institute verification reported here.

Does the lutetium clock meet the authors' criteria?

Semantics aside, we believe there are issues with the application of these criteria to the present work.

First, as applied to lutetium the criteria are logically inconsistent. By insisting that clock accuracy budgets are only "acceptable" insofar as they have been tested with same-species comparisons, the authors imply that rigorous evaluation via secondary measurements is insufficient or "unacceptable" – even where no specific reason exists to doubt such an evaluation. The argument is that to be trustworthy, these evaluations require additional information from a same-species comparison to corroborate the standard uncertainty evaluation. At core, this is an epistemological stance advocating skepticism towards a specific category of scientific results. For this stance to be meaningful, it must be applied uniformly. If only applied where convenient to the authors, it ceases to be rigorous.

The same-species comparison reported here provides no additional information to corroborate the evaluations of the blackbody shift or second-order Zeeman shift, since those are both common-mode between the clocks. Therefore, those aspects of the clock evaluation don't satisfy this epistemological criterion and should not be considered "acceptable".

In response to this point from our initial review, the authors claim that "both of these properties have been rigorously assessed". This is true; the scientific work in this manuscript is of high quality. However, systematic effect evaluations by many other high-quality research groups around the world also feature "rigorously assessed" properties with equally high quality. Insofar as the authors wish to discount those other accuracy evaluations because they haven't been tested in same-species comparisons, the exact same logic must apply to these shifts on lutetium. To be "acceptable", they must be corroborated by additional information from a same-species comparison. Applying this standard to an overall clock evaluation but doing so in a way that does not corroborate the two largest shifts is logically inconsistent.

We note that a same-species comparison at significantly different temperatures could, in principle, uncover errors in calculated or measured atomic properties; in that case, the incorrect BBR shift corrections applied for the two different temperatures would result in a non-unity ratio.

In this light, no clock evaluation can be "rigorous" under the authors' criteria until a same-species comparison has been made between two apparatus at significantly different temperatures and with significantly different magnetic fields. That this would be difficult is not sufficient reason to make an exception just for lutetium or just for the BBR shift; it's evidence that "validating" a clock without relying on rigorous characterization of effects through secondary measurements is, in fact, difficult.

We emphasize that we don't consider this a reason to doubt the evaluation of the BBR shift in this work. Rather, this undermines the authors' asserted criteria for clock evaluations and their criticism of standard practices in the clock field.

Second, the role of the same-species comparison in making an accuracy evaluation more rigorous is to constrain shifts due to unknown effects and due to experimenter error. A self-comparison can constrain these only insofar as they are not common-mode between the two clocks. However, the independence of these two systems is limited in ways that significantly reduce their ability to constrain unknown effects, even beyond the temperature and magnetic field servos. Shared environments, control software, data analysis pipelines, apparatus design, and other operating conditions are all avenues that unknown effects could come in through. To our knowledge, the majority of errors found in previous uncertainty evaluations would not have been caught by a same-species comparison similar to the one performed here; for example, the incorrect atomic structure coefficients or clock laser spectral impurities discussed above would both cancel in common mode.

One of the reasons the SI redefinition roadmap demands different institutions make comparisons is to control for these unknown effects. So-called "dark" uncertainty is unfortunately common in high-precision interspecies clock comparisons, both within comparison campaigns and between measurements of the same ratios (cf. references in section 6C of Fortier et al., Optica 13 143-163 (2026), as well as Aeppli et al., arXiv:2512.21428 (2026)). Same-species comparisons involving independent systems can help resolve these issues precisely by constraining unknown shifts affecting one or more clocks.

The fact that 'dark' uncertainty exists and is common means a significant number of errors in reported uncertainty evaluations exist and the source(s) are unknown. One can't uncover errors in evaluation from same-species comparisons unless they have been carried out. Table 2 of the manuscript tells the story on this without explicit elaboration here or in the manuscript. For interspecies comparisons, one cannot even discover a problem without two independent comparisons. Even then, which comparison was correct, if any, would be speculative without evidence-based identification of the cause.

Third, this work doesn't meet the authors' own "minimum criteria for a rigorous assessment". For one, they are making a "performance claim" at the 1e-19 level but only do a comparison at the 5e-19 level. In the authors' language, the "claimed accuracy cannot be better than the measurement precision" for the same-species comparison. If the authors were reviewing a different group's clock accuracy evaluation, one assumes that they would implement this standard by refusing to accept an accuracy claim with precision beyond the relevant self-comparison. The same should be true here - if this performance claim is "unacceptable", it should be moderated to match this standard or the standard should be rethought.

The intent of the criteria (ii) was, and is, that the 'performance claim' is, in the case of this report, 5×10^{-19} which is limited by the measurement precision. It has been reworded to a more concise "(ii) a same-species comparison to demonstrate validity of the uncertainty budget to the degree the measurement precision allows".

'Claims' are now consistently stated in the abstract, intro, and discussion. E.g. ... lines [57-59] '...each with evaluated uncertainty near 1×10^{-19}...verified at the level of 5.7×10^{-19}...'.
While we no longer imply what is or is not 'acceptable' practice for clock assessment, the principle of (ii) is quite simple: without comparison an uncertainty evaluation is untested, and with one it is tested up to level which precision allows.

For another case, the new criteria require "a stress test on any systematic that is more than an order of magnitude larger than the accuracy claim". As far as we can tell, the authors have not done sufficient "stress tests" on the two clocks presented here. Again, this would require changing the BBR environment significantly enough to evaluate the BBR shift. While the authors argue the differential BBR shift is small for a reasonable range of laboratory temperatures, what needs to be epistemologically confirmed and "stress tested" isn't the differential shift between the two clocks; it's the total BBR shift from 0K to room temperature, which is more than an order of magnitude larger than the claimed accuracy.

Again the 'performance claim' in the previous manuscript refers to 5.7×10^{-19}... under reasonable interpretation of the criteria language. Nonetheless (iii) has been changed to '(iii) a stress test on any systematic that is substantially larger than the measurement precision of the comparison...' regardless of whether the precision is more or less than the unevaluated uncertainty, the measurement precision is the relevant consideration if stress testing is feasible. The total BBR shift is only a factor of two larger than our precision, and in the revised discussion paragraph, lines [258-275], we explicitly discuss our considerations for 'stress testing' in our systems.

The measurements using different fields on Lu-1 (presented in Fig. 3) would satisfy this requirement for Lu-1 on the Zeeman shift, although it's not clear whether that applies to Lu-2 as well – that depends on whether the authors intend this to be a requirement only on a species-by-species basis, or separately for each clock evaluated.

The criteria are given as concise guidelines to be applied with a basic level of common sense. The uncertainty in the Zeeman shift is dominated by an atomic property. Atomic properties do not need to be measured on a per apparatus basis, but should be corroborated by multiple measurements, as we have done in different apparatuses over time, and ideally by multiple groups.

Are these new criteria for "acceptable" clock comparisons a good idea in the first place?

The authors would like to add same-species comparisons to mandatory condition I.1, when they are already present in a more restrictive form in I.2. Aside from this being insufficient to demonstrate "validation" as defined in criterion I.2, we disagree that a same-species comparison should be required in order to report clock accuracy evaluations. We believe this stems from a different understanding of what an accuracy evaluation is or should be. As interpreted in nearly every other published optical clock result to date, the accuracy evaluation is nothing more nor less than a collection of measurements which rigorously estimate shifts and uncertainties to an atomic clock frequency. In other words, the authors' criterion i.

In essence: an accuracy evaluation is not a "hypothesis", as the authors put it in the revised introduction. It is a specific type of reported research result.

This has been changed to, lines [44-46] "An evaluated uncertainty budget is a quantitatively testable prediction that two frequency standards of the same species, corrected for their evaluated systematic shifts, should agree to within their combined uncertainties." The intended meaning was simply that an uncertainty evaluation can and should be tested. We agree that insofar as 'hypothesis' is defined as 'a tentative, testable explanation for a phenomenon' it was not a precise word choice.

An accuracy budget does not amount to a guarantee of the variation in frequency ratio measurements from a "true" value – that includes the significant additional experimental effort which goes into making a high-precision comparison, including frequency transfer, phase-noise compensation, operation of optical frequency combs for inter-species comparisons, relative geopotential measurements between clocks, and data analysis. All of these are opportunities for error which cannot be appropriately included in any single clock's accuracy evaluation.

Adopting the authors' criteria would amount to invalidation of every previously published accuracy evaluation except one, which features identified flaws. Additionally, it would push institutions to redirect resources towards replicating identical apparatus rather than developing new clocks or advancing clock performance, and it would exclude groups without the resources to build two identical clocks.

Moreover, any additional utility from this definition is already captured by the existence of mandatory criterion I.2 – as part of the SI redefinition process, clock error budgets must be validated at 5E-18 by multi-institute clock comparisons. Beyond the redefinition, in the case of the optical second, the CCTF Working Group on Primary and Secondary Frequency Standards would continue to stringently evaluate individual frequency standards at the relevant level of precision, just as they do now for both microwave and optical standards at the E-16 level relevant to TAI.

We agree with the authors that same-species comparisons have additional value beyond the accuracy evaluation. They represent a check against many - but not all - potential significant errors in clock construction and evaluation. They also demonstrate the performance of a clock in a more direct way. Hence, again, why we support publication of the underlying results here in Nature rather than a less prestigious subject-matter-specific journal. However, this utility is additional to the accuracy evaluation, not inherent to it.

While we clearly have different ideas on what are the likely sources of the $\sim 5 \times 10^{-17}$ dark uncertainty in global ratios and how it should best be resolved, this is beyond scope of the revised paper and so we won't debate in this peer-review discussion.

The manuscript is now presented simply as a clear statement of the standard to which we hold ourselves, and clarifies how we interpret the results. We leave it to the community to decide what they wish to take from it. At a minimum, we can hopefully agree that more same-species comparisons, and especially side-by-side comparisons in light of geoid uncertainty limitations, are currently needed at the highest levels of accuracy and precision now reached.

Magnetic field correlation analysis

We agree that the magnitude of the magnetic field noise will not be a significant issue, and the added details make this claim more convincing. Our comments and questions have been addressed. However, the newly added Fig. S2d requires some additional details. First, there is no metric provided for judging the magnetic field noise as 'uncorrelated.' One could claim a small anti-correlation based on visual inspection alone — what is the correlation coefficient? What does the thick black line represent? Second, the quantum projection noise, which is independent between Lu-1 and Lu-2, can dominate the correlation of the magnetic field noise. For the experimental conditions in Fig. S3a, the averaging time would need to be around 100 seconds to remove the projection noise component.

The black line is a linear fit to the scatter data, Fig. S2d, for which the slope is nearly zero. In the revised manuscript this is now quantified by the Pearson coefficient, which evaluates to $r = -0.009$. This data is only indicative that the magnetic field fluctuations are uncorrelated on longer time scales than the Ramsey interrogation, as already discussed in the previous reply, and as noted by the referee is largely immaterial to the results.

For the shorter time scale correlations, as in the previous reply and the previously submitted manuscript (lines 144-152 in the new manuscript): 'Independent measurements on magnetic sensitive states indicate the ambient magnetic field has flicker instability of 7 nT on the timescales of interest for clock interrogation ($\sim$ 1 - 100 s). We find the loss of contrast observed is consistent with simulations of uncorrelated magnetic field noise at this level for each ion, shown as the solid line in Fig. 2a. We note that Ramsey spectroscopy coherence times on the microwave transitions of the individual systems are approximately twice as long as for HA-HR correlation spectroscopy, and both decrease linearly with increased dc magnetic field, which is all consistent with decoherence due to uncorrelated magnetic field noise at the observed level.'

The comment at the end of supplemental has been revised to "We observe no correlations in the magnetic field corrections between the two ions (Fig. S2d), indicating that the magnetic field drifts at the two ions on slower timescales than the servo are uncorrelated. As noted in the manuscript, Ramsey coherence times on the microwave transitions in the individual traps relative to the contrast loss in correlation spectroscopy indicates that magnetic dephasing is also uncorrelated between the two systems." to reinforce this point.

SODS and heating rate measurements

We were disappointed to learn that the authors had not measured the secular motional energy in one of the three secular modes for both clocks. The authors are correct about the distance scaling of anomalous heating, and that heating rates typically scale smoothly with frequency, but both significantly anisotropic heating and anomalous frequency scaling have been reported in other macroscopic traps. For example, Dawel et al (arXiv:2509.22525, 2025) observed a factor of 5 larger heating rate along one radial direction than the other, using a blade trap generally similar to that used in this work and with a difference in secular frequencies of less than 0.5% between modes. Lindvall et al (PR Applied 24, 044082, 2025) observed heating rates that changed by a factor of 5 over a measurement campaign, and depended on trap secular frequency beyond a smooth scaling. Leopold et al (Rev. Sci. Instrum. 90, 073201, 2019) observed a factor of 2 difference in heating rates for a cryogenic blade trap with a $\sim$10% difference in mode frequencies.

We note that while the trap used in this work demonstrates mechanical symmetry, there is not a matching electrical symmetry - according to ref. 50 of the revised manuscript, one pair of rod electrodes provides the RF pseudopotential and the other pair is held at DC. This implies different impedances to ground and different electrical sources connected to the pairs, which is a natural way anisotropic technical noise could affect the ion.

This is an area where the same-species comparison is valuable - realistic sources of anisotropic heating are likely technical in nature rather than inherent to the trap design, so they're unlikely to be correlated between the two systems. However, this only constrains the effect at the 5.7E-19 level, rather than the 1E-19 overall accuracy. Perhaps the authors can derive a more stringent constraint from the dephasing measurements, or measure the heating rate of the inaccessible mode using parametric mode coupling as demonstrated in Gorman et al (Phys. Rev. A 89, 062332, 2014). However, it's not obvious to us, given these results from a brief literature search, that potential anisotropy can be accounted for simply by increasing the uncertainty on the unmeasured mode from ∼5% to ∼10%.

At the behest of the reviewer, measurements were made along the other radial mode. Since these measurements were made a year after the fact, two sets of measurements were made 1 week apart to check for consistency. New measurements on the previously measured modes were all statistically consistent with previous results indicating long term stability. The two measurements for the previously unmeasured mode in Lu-2 were statistically consistent giving 70.4(5.5) μK/s, which is also consistent with expectations from the other mode. For Lu-1, new measurements for the previously unmeasured mode were also statistically consistent but the value 110(15) is roughly a factor of 4 higher than the other mode. From measurements at different drive powers, the higher heating does not appear to be near to a region of instability as observed in Lindvall et al (PR Applied 24, 044082, 2025). However, the rate is approximately consistent with those of Lu-2. Since the mode with the significantly lower heating rate is along the axis determined by the rf drive and the resonator Q for Lu-1 is significantly higher, we would speculate that any noise into the rf-resonator is better filtered for this mode based on comparison of the radial trapping frequencies with the bandwidth of the respective resonators.

The section has also been updated to account for a previous minor oversight: by using the approximation $\bar{n}_i \approx k_b T/\hbar\omega_i$ this neglected the SODS from zero-point fluctuations. This is only a 6% correction to the SODS shift for Lu-1 but included for completeness.

The updated heating rates and SODS evaluation shifts the measured frequency difference from -1.5×10^{-19} to -0.1×10^{-19}, well within statistical precision, and with total systematic uncertainties unchanged.

First-order Doppler shift

We agree with the authors' discussion of charging and discharging from UV photoelectrons. This is an unrealistic possibility when applied to the Lutetium case. The point of bringing up the NIST group's measurement was not to suggest a particular source of the FODS, but rather to make the point that large first-order Doppler shifts can arise in ion trap systems and be difficult to detect without counterpropagating beams. The authors point out an issue with the photoelectron hypothesis from that experiment - if they are correct, that would only further emphasize that first-order Doppler shifts can arise without an obvious mechanism, and should be constrained by measurement.

Photoelectrons aside, the authors have three responses to our concerns about the potential first-order Doppler shift, which we will summarize here for clarity of discussion.

1. The self-comparison constrains any FODS, because the odds that it would be identical and synchronized between clocks are vanishingly small
2. The authors dispute the likelihood of the specific possible FODS scenarios we discussed in our initial report.
3. Constraining the FODS with new measurements using counterpropagating beams would take too long to average down, and "incorporating a purely hypothetical FODS into the uncertainty budget lacks scientific utility"

We will address each of these points in turn. Regarding the first point - we agree that the odds of an undetected FODS are small, and it is reasonable to constrain it at the level of the self-comparison. However, this comparison is five times less precise than the claimed systematic uncertainty, so this point is only partially relevant. This is related to the contradiction discussed above between the authors' performance claims and their epistemological stance on the necessity of self-comparisons.

Regarding the second point - we initially note that all of these proposed mechanisms were intended to illustrate the difficulty of constraining an FODS from first principles, rather than detailed estimates. In line with the authors' epistemological arguments, first-principles arguments such as these are relatively unsatisfying. A few specific notes:

- We agree that scattering from the ion is negligible. To our knowledge, the MW power used is not given in the manuscript; at single-digit milliwatts, we agree that is also unlikely to cause substantial differential heating across the trap.
- The optical powers used for interrogation were similarly not given in the manuscript. If the figure of 10 μJ per clock cycle includes laser cooling and readout as well as clock interrogation, then we agree the total power is likely too small to cause a substantial FODS. We appreciate the authors' provision of additional details.
- Regarding Lorentz forces, the concern is not oscillating currents, which as the authors point out average to zero for currents at different frequencies, but rather DC forces from power supply currents and ground return currents, which may create a different electric-field environment when the laser or microwaves are off vs on. This is also likely too small to matter but again requires microscopic knowledge of the experimental apparatus, e.g. where do ground currents flow / how well have ground loop currents through the vacuum chamber been eliminated, etc.

- The timescale argument is only partially convincing. For example, suppose there is a mechanical vibration on the milliseconds scale from a shutter closing. To average 30 pm/s over the seconds-long probe time, the scale of the vibration at the ion or at one of the clock laser beam-path mirrors would have to be 30 nm/s, which still does not seem unreasonable if the shutter is close by and depending on the spring tension in the relevant mirror mounts.

- Relatedly, we note that ion clock experiments have observed first-order Doppler shifts related to the transient response of path-length servos to mechanical-shutter-induced vibrations. We believe this is already constrained by the AOM phase chirp measurement, but mention it to point out that the timescale question can be quite complicated - in this case, the transient response extends the timescale of the perturbation significantly beyond the timescale of the mechanical vibration.

The reviewers have revised their objection with suggestions of switching currents "changing the electromagnetic environment of the ion" through ground loops, and the possibility of mirrors moving, which they claim can happen even with path-length stabilization in place.

1. Switching currents: The initial mechanism suggested was Lorentz forces which do not act on a stationary charge. So the only mechanism would be strain on the ion-trap structure. Any reasonable ground loop current would be much less than the rf currents induced by the high voltage driving the trap (a capacitor) and nowhere near the same proximity as the two rf electrodes. As we already argued, any force between the rf rods will not conceivably give significant movement.

2. Switching currents: The reviewer has now changed the narrative from Lorentz force to "changing the electromagnetic environment of the ion". If the proposed mechanism is acting via Zeeman shifts then this is also not feasible for many of the same reasons. It cannot act over the entire interrogation time and any such effect is heavily suppressed. In addition, we have looked for such effects when doing direct microwave spectroscopy and saw no evidence to support such an effect (as already stated in a previous response).

3. The moving mirror hypothesis: As path lengths are actively stabilized, this is unrealistic. We acknowledge the possibility of transients due to a dynamic change, but 30 nm is a very large deviation in an interferometer that would be rapidly captured by a functioning servo. We would also acknowledge the possibility of an *observed* large FODS in the presence of path-length stabilization but only as an indicator that the path-length stabilization was not functioning properly. In our case, that would also have to be mysteriously common-mode to the two systems, i.e. not functioning in the same way for two independent implementations, for it not to appear in the comparison.

The moving mirror hypothesis highlights an internal inconsistency in the reviewers position. For the sake of argument, suppose such a situation exists. Take the beam from one apparatus and direct it into the other in a counter-propagating configuration as done by the NIST group [PRL 123, 033201 (2019)]. Whether the movement is common-mode or not, shifts that arise would not cancel in an average of counter-propagating probes as they are not due to the ion moving. This clearly is not consistent with cancellation claimed in [PRL 123, 033201 (2019)]. If the mirror movement is common-mode, as it would have to be to not be detected in our comparison, it would also not be seen in a counter-propagating configuration. If it is not common-mode it could only ever be bounded by the measurement precision as is the case for our comparison, and as is the case for any experiment ever carried out for such hypotheticals that defy reasonable, testable, investigation.

Regarding the authors' third point - we fail to see a meaningful difference between incorporating a "purely hypothetical" error source into an accuracy budget just because it isn't experimentally constrained, and rejecting rigorously evaluated uncertainty budgets that don't come with a self-comparison to reject other purely hypothetical effects.

It is not that the purely hypothetical error sources are not experimentally constrained. It is that they have no plausible scientific rationale or require unrealistic assumptions to explain the experimentally observed agreement. The moving mirror hypothesis clearly illustrates the unending and ultimately circular nature of such hypotheticals.

That aside - while it may be unsatisfying, we acknowledge that the standard practice in the field is to omit the FODS from uncertainty budgets unless there is specific reason to include it. We believe this leaves unanswered the question of whether large FODSes of unknown origin, similar to that observed by the NIST group referenced earlier, may also be present on other ion clocks. We believe this work would be better-supported had the authors conducted their campaign with counterpropagating clock beams. However, we acknowledge that they are reasonable at the level of standard practice in the field, especially when combined with the additional information provided in their response but not included in the manuscript, and don't require any additional changes to the manuscript on this front. We appreciate the inclusion of bounds on some FODS sources that can be estimated from the data the authors did collect.

We thank the referee for the discussion and note their conclusion that no further manuscript changes are required.

That said, there is a certain tension with the authors' arguments about what constitutes an "acceptable" or "unacceptable" accuracy claim, where they discount rigorous evaluation and first-principles arguments as unverified without a self-comparison. Are the authors truly satisfied with this evaluation of the first-order Doppler shift, and if so what differentiates it from other clocks' "rigorous evaluation" of systematic effects which the authors consider "unacceptable?" The measurement of the first-order Doppler shift using counterpropagating clock beams would constrain otherwise un-measureable effects in the same way that a self-comparison would constrain some unknown differential effects; however, the authors decided not to make this measurement, while simultaneously claiming that accuracy budgets are not acceptable with uncertainty below the stability of a self-comparison.

As noted above, a large FODS would be seen in our comparison as there is no good reason it would be a common-mode shift. We note the only other high precision correlation spectroscopy comparison, carried out by the NIST group [Clements PRL 125,243602 2020], did see large frequency differences that were not consistent with their UV charging hypothesis. Although the frequency differences are not explicitly reported in that work, they can be inferred from the phase offsets used in their fits and the interrogation times ($\phi = 2\pi\Delta f T$). The minimum possible shifts estimated from this are larger than that which can be explained by Zeeman shifts from their stated field stability and do not decay with interrogation time, as it would under the UV charging/discharging hypothesis.

We also note that a new trap from NIST (PRL 135, 033201 (2025)) and results from PTB (arXiv 2606.23151v2) show no evidence of a FODS for an Al+ clock. So we have gone from a clear 1e-16 difference to no evidence of a difference at the low 1e-18 level with no explanation of design considerations to explain this dramatic improvement. Moreover, new ratio measurements with Al+ (PTB arXiv:2606.23151 and BACON2 arXiv:2512.21428) increasingly indicate substantial disagreement with those obtained with the trap having the "UV charging" problem, for which counter-propagating beams supposedly suppressed.

The onus is not on us to investigate problems for which there is no plausible scientific mechanism for them to arise, and any evidence showing they don't exist can be rebutted by a new and equally unreasonable hypothetical. When any group is faced with undeniable, observable evidence of a problem — as in the NIST case — the onus is on them to investigate its origin and deliver a scientifically rigorous, evidence-backed explanation.

If the authors want to present their clock as not just a significant experimental achievement, but as a model for scientific rigor and example of a framework that other atomic clocks should follow, they must also address the places where their work is in tension with that framework. Similarly to the lack of "stress tests" discussed earlier in our review, the fact that many potential sources of FODS are constrained only by first-principles argument rather than measurement should be addressed, or else the framework and epistemological arguments should be modified to be consistent with the experiment.

This is a misinterpretation of the presented framework. Lines [258-275] now discuss the meaning and application 'stress tests' to our system, which may be helpful should others choose to apply this framework to their systems. For the wider readership, it also serves to highlight important considerations for critically evaluating the performance claims of clocks in general.

Small comments

Minor typo on page 2 of supplemental material: "We take this power to be half the total rf drive power assuming equal dissipation at each end of the trap **and**, i.e. all the power is dissipated in the ceramic."

Fixed

Regarding the BBR shift: We appreciate the additional detail included in the new supplemental material. The authors have adequately supported their estimates of the BBR shift.

Similarly, we find that the questions about the magnetic field servo have been addressed.

Response Reviewer #4

I appreciate the authors' efforts to revise the manuscript and their detailed responses. The work remains technically strong and addresses an important problem in optical frequency metrology. However, several of the central concerns raised in the initial review are only partially addressed, and in some cases the revisions remain difficult to assess based on the response provided.

Accessibility and clarity for a broad readership The authors indicate that the manuscript has been substantially revised for accessibility, including moving technical material to the Methods and adding explanatory text. However, the response does not identify specific changes or provide representative examples demonstrating how key concepts have been clarified.

Given the importance of accessibility for Nature's readership, it is important that central ideas such as correlation spectroscopy, hyperfine averaging, and the structure of the interrogation sequence are introduced with clear conceptual explanations prior to their quantitative use. Based on comparison with the original manuscript, the revisions appear to consist primarily of minor rewording and relocation of technical material, rather than the addition of conceptual explanations needed for accessibility to a broad readership. Ultimately, the revisions are difficult to verify and appear limited in scope.

As requested we have introduced additional conceptual explanations to improve accessibility:

- Correlation spectroscopy (lines [75–77]): "... a powerful technique to directly measure the phase evolution of one ion relative to the other, rejecting common-mode phase noise from a shared clock laser, and enabling interrogation times limited only by the atoms..."

- Hyperfine averaging and interrogation sequence (lines [100–117]): Expanded paragraph to conceptually introduce Hyperfine averaging, Hyper Ramsey, and explain the utility offered by this interrogation sequence, which is fairly unique to our system, over other possibilities.

Demonstrated, inferred, and projected performance The distinction between experimentally demonstrated performance and inferred or projected capability remains insufficiently explicit. While the authors have moderated some language, projections of single-clock instability based on correlation spectroscopy measurements rely on assumptions (e.g., availability of state-of-the-art laser stabilization and comparable interrogation times) that are not realized in the present work. These projections should be clearly identified as conditional and separated from the primary experimental results.

Projections of single-clock instability have been removed from the introduction and now the introduction paragraph, lines [72-85], is restricted to comparison measurement instability between independent systems involving at least one ion clock. This simplifies the unambiguous and direct comparison of our demonstrated measurement instability as a figure of merit to compare to other cited reports.

Both inferred and projected stand-alone clock stability, and remaining forward-looking statements (operation at 25 μT, interrogation times up to 30 s), are confined to the Discussion outlook and are explicitly framed as conditional (lines 288–309]).

Interpretation of performance under shared-laser interrogation The manuscript would benefit from clearer framing of what aspect of "clock performance" is being probed. Correlation spectroscopy provides a powerful measure of atomic reference agreement under shared-laser interrogation; however, for a broad readership, this is not equivalent to the performance of a fully independent clock system.

A more explicit distinction between 1) atomic reference consistency under common-mode rejection, and 2) standalone clock performance would help avoid potential misinterpretation of the results.

The distinction the referee requests is now stated explicitly (lines [78-80]): "The measured agreement therefore characterizes the atomic references themselves, independent of the performance of any particular clock laser." Consistent with this, the title and text now use "optical frequency references" rather than "clocks" when referring to the compared systems.

Framing of "validation" and statistical resolution The revised manuscript introduces a more explicit definition of "validation," which is helpful. However, the central issue remains that the measured agreement is limited by statistical uncertainty at the 5 x 10-19 level and therefore does not resolve agreement at the 10-19 level. It is important that this distinction is consistently and unambiguously reflected in the manuscript, so that claims of performance at the 10-19 level are clearly framed as inferred or consistency-based rather than experimentally resolved.

The framing of the discussion in term of the roadmap and validation has been removed. The wording of the 'criteria' tightened and now presented only as a framework for our clock assessment. The manuscript now consistently separates the two figures: the evaluated uncertainty (1×10^{-19}, inferred from auxiliary measurements as standard practice in the field) and the 'verified' performance (5×10^{-19}, limited to the level resolved by direct comparison) in the abstract (lines 26-28), introduction (lines 56-59), and discussion (lines 230-232).

In summary, while the manuscript has been improved, further clarification is needed to ensure:

- accessibility to a broad readership,
- clear separation of demonstrated and inferred performance, and
- unambiguous framing of what aspect of clock performance is being measured.

Addressing these points would significantly strengthen the manuscript and clarify its suitability for Nature. As written, the technical specificity to those very familiar with optical metrology and optical atomic clocks makes of the manuscript more suitable for PRX.

Referees report on 2025-12-32495

January 28, 2026

1 Overview

The manuscript reports on two optical frequency references with systematic uncertainties of about 1×10^{-19} and demonstrates a frequency difference measurement at 5×10^{-19}, limited by statistical uncertainty. This is the first full evaluation of the Lu^+ clocks, after years of efforts by the authors, and establishes Lu^+ as an excellent frequency reference. This work sets two world records, for clock accuracy and clock comparison precision. Moreover, each is a significant improvement - the systematic uncertainty is five times lower than the existing record, and the comparison uncertainty is twice as low as the previous best. The combination of these two achievements is more than the sum of its parts - the comparison helps to justify or build confidence in the accuracy, a step beyond previous record-setting clock evaluations. At a time when the global scientific community is undertaking a major effort to redefine the SI second, we believe this manuscript is appropriate for the wide readership of Nature.

The manuscript is largely valid, with careful evaluations of many potential sources of systematic error. The systematic evaluation as it stands currently is incomplete, as we discuss below; but we are confident the authors can address our concerns in revision, either with additional measurements or further analysis. The manuscript is presented efficiently, but contains sufficient information for the expert reader to fully evaluate the results. Statistics are evaluated appropriately using standard tools in the field.

In sum, we believe this manuscript merits eventual publication in Nature.

However, before publication we have several concerns that must be addressed. Two significant systematic shifts need additional attention - the first order Doppler shift and the evaluation of the effective blackbody temperature. See below for detailed discussion of these, as well as for comments and concerns about the rest of the manuscript.

Before the specific elements, however we wish to raise a general point which recurs across the manuscript. This concerns the discussion of “validation” and “independent” clocks. We agree with the authors that same-species comparisons are essential to progress towards the optical second. The work presented here does represent the most accurate same-species comparison to date and is a significant step for ion clocks and for optical clocks generally; the inclusion of a same-species comparison together with the uncertainty budget is what brings this work up to a higher level.

That said, we find the authors' discussion of the same-species comparison as an independent validation of the uncertainty budget to be slightly overstated. These two clocks are independent in that they have separate vacuum chambers, but their systematic effect evaluations still overlap to a significant degree. They are servoed to the same magnetic field and environmental temperature; share the same laser and RF sources; are operated with synchronized clock cycles (meaning they share laser noise, environmental noise, etc); are operated under a single differential servo algorithm; and are run, designed, and analyzed by the same scientists. The comparison validates the evaluation of differential systematic shifts, such as AC Zeeman shifts, second-order Doppler shifts, etc. However, the two largest systematic shifts – quadratic DC Zeeman and blackbody radiation – are essentially common-mode. Similarly, the comparison does not bound unknown common-mode systematic effects. We find no issues with how these effects are evaluated - the authors have done careful and thorough work in this and other papers to measure these effects. The point is that the same-species comparison does not make these parts of the evaluation any more "valid" than a careful evaluation with a single clock, as errors with their evaluation would cancel out of the comparison.

The authors should address these limits, whereas the manuscript in its current form could be interpreted to claim that the comparison completely validates the full uncertainty evaluation. For example, the authors could clarify what is meant by "independent" in the context of the two Lu+ clocks by adding a sentence in the "discussion" section, perhaps in the context of the existing discussion on line 207 of the degree to which perturbations are uncorrelated; or alternatively use more specific language when describing how the two clocks are independent on the first page; etc. Additionally, the claim on lines 75-76 of "strong evidence that all systematic shifts above the 5×10^{-19} level have been accounted for" should be moderated - this is only true for differential systematic effects between the two clocks.

We also point out that the term 'validation' is more rigorous in the roadmap defined from the international metrology community (I.2 category, ref. 2). At least three experiments (refs. 6, 7, 15) have shown unity frequency measurements below the 5×10^{-18} criterion. Although these works have similar or greater independence compared to this work, the community does not consider them as 'validation' since the comparisons were performed within a single laboratory. Given this apparent consensus on the term 'validation' within the metrology community, and given that the authors also take such consensus seriously, the claim of validation should not be misleading. While the authors discuss this more stringent interpretation of the term 'validation' in the Discussion section, given this difference in definition and the fact that only some effects are addressed by the comparison, we question whether the term "validated" is justified in the article's title.

In the following, we first give comments on the main text and then go through each systematic shift. The two systematic shifts we mentioned earlier are discussed in the 'First order Doppler shift' and 'BBR assessment' section.

2 Main text

On line 21, 'difference' instead of "agreement"?

On lines 66-72 the authors state that the potential instability of a single lutetium clock, with an ideal noiseless clock laser, of 2.4×10^{-16} at one second is an improvement over the lowest instability demonstrated by a trapped-ion clock. This claim is somewhat misleading - a better apples-to-apples comparison would be with the manuscript's reference 27, where a correlation spectroscopy stability of 2.8×10^{-16} was achieved. The same analysis would imply a laser-noise-free limit of 1.4×10^{-16} at one second.

On lines 75-76, while the statistical uncertainty of the comparison is 5×10^{-19} and the measurement is consistent with zero, it's not obvious that 5×10^{-19} is the right level to claim justification of the differential uncertainty evaluation. The $1 - \sigma$ confidence interval is (-7.8,+3)$\times 10^{-19}$, so errors in the evaluation of differential systematic effects are only excluded with 67% confidence below 7.8×10^{-19}.

Regarding Fig. 1 - is AOM 1a used to cancel common-mode path length fluctuations between the ions and the shared clock laser source? Also, delivering the maser signal accurately at 3×10^{-15} is also not so trivial. Could you describe how the setup is designed?

Regarding Tab. 1, the background gas shift is 0. Is this because of the collision detection algorithm described in Methods?

On lines 82-83, the authors mention two central methods, hyperfine-average and hyper-Ramsey without any citation or explanation. It will be useful to have citations and brief descriptions for the broad scope of readers (what is averaging and not averaging in HA, probe Stark is canceled in HR). Besides, the pulse sequence seems to be slightly different from [16].

On line 84, $\sigma_{z,i}$ is not defined.

On line 85 - We couldn't find a definition of $f_{6,7,8}$ anywhere in the text or figures. If ν is used for representing atomic frequency and f for the field frequency, $\nu_{6,7,8}$ can be used with the definition.

On line 87, is $f^{(k)} = f_L + \frac{1}{3}(2f_1 + f_2)$ or $f^{(k)} = f_L^{(k)} + \frac{1}{3}(2f_1 + f_2)$? The current form gives $\delta f = 0$ always.

Regarding Tab. 2, there are several points that should be addressed. The systematic uncertainty of the reference 15 should be 1.4 instead of 6. The 6 contribution is reported as the gravitational redshift from the Terrestrial Time surface, which can be neglected in local measurement like this work. A recent work from VTT MIKES could also be included in this table [T. Lindvall et al., Phys. Rev. Applied 24, 044082 (2025)]. The authors include their previous work with a total systematic uncertainty. We could not find the explanation of both systematic uncertainty and the comparison uncertainty. In the reference 16, the authors "refrain from making any accuracy claim" because of the lack of a BBR assessment. Finally, it would be useful to include the mean values of the comparisons in the last column of the table, for the similar reason we mentioned earlier.

On lines 109-113, the authors report 7 nT flicker (1/f) instability of the magnetic field for a 1-100 s range. The loss of contrast is also attributed to this noise. Although the HA-HR protocol largely suppress the quadratic Zeeman effect from the hyperfine splitting, depending on the interrogation scheme, there may still be residual sensitivity to the time-correlated noise such as oscillations at the clock cycle period or 1/f noise. In [T. Lindvall et al., Metrologia 60, 045008 (2023)], it was reported that the servo error can be as large as several parts in 10^{-18} for about 10 pT flicker noise environment with about 100 kHz Zeeman splitting. Here, we have 3 orders of magnitude larger magnetic field noise but similar or larger suppression of field sensitivity (about 20 Hz at 0.1 mT?). How significant could this uncertainty be? It will be common-mode between Lu-1 and Lu-2 but we think it should be included in the overall systematic uncertainty if it is not negligible.

Regarding Fig. 2, we could not find how the error bars in Fig. 2c is estimated.

On line 218, the authors predict the magnetic field noise would be reduced if they use a smaller magnetic field. Is the magnetic field noise dominated by the current driver of the coils?

3 First order Doppler shift

The most significant gap in this work is the omission of the first-order Doppler shift (FODS) systematic. This has the potential to be a significant systematic shift in any clock, and especially in trapped-ion clocks; for example, in the manuscript's reference 37, the FODS was measured to be 4.6×10^{-17}, about 460 times larger than the total uncertainty presented here.

A FODS is induced by any relative motion between the ion and the laser beam path (more specifically, the retroreflecting mirror used as phase reference, see manuscript Fig. 1a) that is synchronized with the clock interrogation cycle. To shift the clock frequency by 1×10^{-19} requires only $10^{-19} * c \approx 3 \times 10^{-11}$ m/s or 0.03 nm/s, less than 10^{-4} of the clock laser wavelength per second.

While the details of the ion trap in ref. 37, which trapped two ions rather than only one, are different from the work presented here, there remain many possible sources of a FODS at the level of 10^{-19}. Some that come to mind include:

- Thermal expansion/contraction in trap or optical materials due to laser or microwave absorption;
- Microwave-induced forces, including electrodynamic forces from eddy currents or ponderomotive or dielectrophoretic forces on trap or vacuum structural materials;
- Lorentz forces due to changing currents in the trap environment e.g. from turning on/off AOMs, changing supply current to MW amplifiers, etc;

- Vibrations induced by shutters or other mechanical elements on the laser table;
- Coulomb forces due to charging of surfaces near the ion, e.g. from thermal or photoelectrons induced by laser beams.

In fairness to the authors, this oversight is not unique to the present work or the Lutetium clock. Several other atomic clock uncertainty evaluations published in the past few years have also neglected discussion of the FODS. However, it is incumbent on the authors to be complete in their uncertainty evaluation, especially when claiming a dramatic improvement in systematic uncertainty over any other atomic clock and especially given the thorough attention to detail in evaluating many other systematic effects.

An order-of-magnitude estimate can illustrate how this is a serious issue which needs to be measured or constrained rather than ignored. Taking thermal expansion of the trap materials as the example, and using the published dimensions of the BeCu trap and endcap rods at $\geq$2 mm long, we see that thermal expansion by 1.5×10^{-8}/s is sufficient motion to give a 10^{-19} FODS. Using average BeCu CTE of 18×10^{-6}/K, this requires a temperature change of 0.8 mK/s (or 4 mK averaged over a 5 second Ramsey time). Using a published rod diameter of 0.45, BeCu density of 8.3 g/cm^3, and BeCu specific heat of 0.4 J/gK, this would require 4 μJ deposited in the rod, or an average of 0.8 μW over the Ramsey time. Energy deposition at this level is not an outlandish possibility; it could come from, for example, backscattered Doppler cooling light, or induced currents from the HA or magnetic field tracking microwave drives. [Note that published trap dimensions for this estimate are taken from Kaewuam et al., J. Mod. Optics 65, 592-601 (2018). Also note that it took clicking through the references of at least five publications on Lu+ clocks to reach this information; the authors should consider adding a more direct reference with trap details to lines 79-80.]

To be clear, this is a rough estimate, not taking into account how the trap electrodes are attached to their support structure, the relative direction of the trap and clock laser, the thermal conductivity out of the trap rods through bias connections (and through spacers, although Macor is a strong thermal insulator), etc. The point is just that the amount of energy required for 0.03 nm/s of ion motion is very small, and could realistically be delivered as thermal expansion or electromagnetic forces.

For a less quantitative example, the authors have shown that dielectric loss in the insulating spacers leads to a large difference in trap temperature between different batches of alumina spacers. Those same spacers are exposed to various microwave fields; while the exposure is shorter and the fields are (likely much) weaker than the trap RF, very small levels of structural heating can cause a FODS. On a macroscopic level, microwave fields will induce currents in the trap vacuum chamber, trap support structure, magnetic field bias coils, optics support posts, etc; these currents will in turn induce heating in and Lorentz forces on various structural elements, again leading to possible motion.

To lead to a FODS, this motion must be synchronized with the clock probe cycle, rather than averaged over. Laser-induced effects would be synchronized in this way, since detection, cooling, and state preparation occur in a block between probes; the HA microwaves are more evenly spread across the Ramsey time, but microwave duty cycle differences between the clock probe and auxiliary measurements also induce a level of synchronization. Effects due primarily to auxiliary measurements or intermittent re-cooling after collisions would be synchronized in the same way.

We note that probing from opposite directions using counterpropagating clock lasers provides a technical solution to the FODS issue, as demonstrated in ref. 37. Alternatively, self-comparisons with different Ramsey times and auxiliary measurement duty cycles will constrain many possible sources of an FODS, but would be limited by the statistical uncertainty of the self-comparison. (The correlation spectroscopy data presented here already constrains any differential FODS to be $\lesssim 7 \times 10^{-19}$.) This effect must be accounted for and incorporated into the uncertainty budget before publication.

4 BBR assessment

The extraordinarily low differential polarizability of the Lu+ clock transition means that the trap temperature evaluation is much less stringent than for any other atomic clock. However, the temperature uncertainty is still relevant and worthy of some attention, given the small total uncertainty. For example, the 1.6 K uncertainty for Lu-2 still constitutes a 4×10^{-20} systematic uncertainty and would be the third largest source of uncertainty overall.

Moreover, it's not entirely clear what the 0.6 (1.6) K uncertainties represent. The lutetium ion doesn't see a single blackbody spectrum, but a combination of blackbody radiation contributions from the trap electrodes, support materials, and vacuum chamber, which emit at a range of different temperatures. With the uncertainty given being equal to the difference between the vacuum chamber and ion trap, it seems likely that it's intended to encompass this full range of blackbody temperatures. This is a valid, well established way to ensure the evaluation is accurate without detailed modeling of the BBR contributions from different surfaces, accounting for their emissivities, etc. However, if the given uncertainties are accounting for this spectral range, then the measurement uncertainty is missing from the evaluation. Alternatively, if the given uncertainties are estimates of the measurement uncertainty, then the spectral range of the blackbody radiation still needs to be accounted for.

The dummy-trap technique used here is in general a valid way to evaluate the blackbody environment. However, there are some additional nuances compared to actually measuring the temperature inside a clock during operation, which introduce additional sources of uncertainty.

1. To make the single measurement of ΔT_k a valid evaluation of the range of blackbody temperatures seen by the ion, it should be measured at the

highest-temperature part of the environment; otherwise, it's not guaranteed that the ±100% uncertainty includes the full range.

1.1. If dielectric loss in the spacer is the largest RF-induced heat source, then this would be the spacers themselves or the endcap rods. However, only the temperature of the end block is measured. Taking the end block as a single temperature should be valid due to the high thermal conductivity of copper, but Alumina and Macor have thermal conductivities that are 15 and 270 times lower (respectively). (See e.g. the manuscript's reference 14, where the trap support structure was thermally anchored but the electrode temperature changed significantly with RF applied). Have the authors modeled or considered gradients in the spacers?

1.2. Additionally, in some ion traps dielectric loss in vacuum feedthroughs can be larger than in spacers, resulting in higher temperatures on the vacuum chamber near the feedthrough. (See discussion in Doležal et al, Metrologia 52 842 (2015).) In others, resistive loss can be higher, whence heat would conduct through to the BeCu trap rods (see e.g. Lindvall et al, PRApplied 24 044082 (2025)). Because the spacers are made from thermally insulating materials, the BeCu trap rods could be significantly hotter than the end block. Were measurements made at these other positions, or is there a way to constrain that the end block does in fact give the highest temperature seen by the ion rather than the trap rods or the RF vacuum feedthrough flange?

2. There is no uncertainty contribution for how well the dummy trap is expected to replicate the real trap. Going based on the drawings in Fig. A3b without additional information, it looks like the physical trap assembly should be very repeatable and not introduce additional uncertainty. However, other possible differences between the dummy and real traps should be considered.

2.1. The loaded Q for the dummy trap resonator is 25% higher for Macor and 40% higher for alumina than the real traps. This would seem to imply more loss in the RF system for the real trap, and therefore a potentially higher temperature. Can the extra dissipation be accounted for? If not, this should be included as a source of uncertainty.

2.2. Does the dummy trap accurately reproduce the RF environment of the real traps, including vacuum feedthroughs and nearby DC/RF electrode bias lines? These could change the distribution of current and therefore temperature in and around the dummy trap.

2.3. The authors observed a large difference in heating between alumina spacers from different manufacturers. What constrains differences between the alumina spacer used in the dummy trap and the real trap - are they known to be machined from the alumina with the same impurity composition? The authors consider the possibility that the original spacers may not have been made from alumina as the manufacturer claimed. If there's a concern about manufacturer malfeasance, is it reliable that these spacers were made from the same batch of material? If additional pieces from the suspect manufacturer are available, this could be tested experimentally. If that is not possible, the assumption that the dummy and real spacers share the same composition should be explicitly stated in the text.

3. The authors measure a temperature gradient across the enclosure. What about temperature gradients across the vacuum chambers? For example, a steel vacuum chamber could have a gradient if the stabilized air temperature is different from the laser table temperature (for example due to thermal conduction across or through the bulk of the table, if the stabilized temperature is different from the surrounding room temperature). If such a gradient does exist, it may be appropriate to add it to the RF heating linearly rather than in quadrature, as it would extend rather than shift the range of blackbody radiation seen by the ion.

4. The calibration uncertainties of the PT100 sensors do not appear to be specifically accounted for. These should be added in quadrature with each other and linearly with the temperature range discussed above (since the temperature range isn't the width of a Gaussian distribution of possible values, but rather a spread of temperatures which all must be accounted for, it can't be treated as uncorrelated with the measurement uncertainty).

The authors should account for these additional sources of uncertainty on the measurement of the temperature, to build more confidence that this evaluation is accurate despite not measuring the actual trap used in the clocks.

5 Magnetic field servo

It seems like a servo error on the magnetic field servo would translate directly into a clock frequency shift. In the data shown in Fig. A1, the out-of-loop measurements constrain the average field value at the 1 nT level, which is a negligible second-order Zeeman shift. However, this only shows one hour of data and servo error can change depending on the behavior of the system being servoed (e.g. drift rate) - were similar out-of-loop measurements taken during the comparison, or can servo error be constrained in some other way? Additionally, we find no mention of the mean value of the difference between the in-loop and out-of-loop magnetic field measurements. Presumably, this implies the difference is smaller than the statistical uncertainty, but this should be stated explicitly.

Additionally, could you explain how in-loop measurement of the magnetic field averages as 1/tau not sqrt(1/tau)?

6 Second-order Doppler shift

A couple of minor questions here. Do the authors have any thoughts as to why the measured temperature is so far different from the calculated Doppler limit, given that the heating rate is quite low? While measurement should be trusted over theory for the trap temperature, an explanation for the discrepancy could help build confidence in the result.

Additionally, were temperatures and heating rates measured on both radial secular modes, and did they agree?

7 Coupling errors

As I understand this section, the coupling error is obtained from a set of three measurements, with one each taken at the beginning, middle, and end of the measurement campaign; and the uncertainty on the coupling systematic is the maximum spread of these three values. Can the measurements distinguish between long-term drift and shot-to-shot or day-to-day variation? Do the authors expect that three measurements will adequately capture the full range of couplings across the measurement campaign?

Additionally, is the uncertainty on each individual measurement negligible compared to the spread between measurements, and if not has it been incorporated into the stated uncertainty on the shift? Since this is the second largest source of uncertainty for Lu-2 (and since it also determines the AC Stark shift, see below), a bit more detail would be useful.

8 AC Stark shift

The uncertainty here appears to be based on the "observed variation in the optical couplings", i.e. the same set of measurements discussed in the "Coupling Errors" section. If so, it may not be appropriate to add these two effects in quadrature in the uncertainty budget. The "Optical coupling" entry for Lu-1 is negligible but for Lu-2 it may be more appropriate to linearly add the uncertainty contributions due to optical coupling errors and uncertainty on Δ_S, and then add in quadrature the contribution from uncertainty on the measured heating rates.

9 Differential AOM chirp

"Differential AOM chirp" section seems to discuss about two different measurements. The first part is on a proxy of the out-of-loop measurement on the clock laser phase, which is closely related to FODS we concerned earlier. However, we could not find the result of this measurement. Although this measurement does not include the sources of FODS on the trap, it would give some idea. What is the result of this measurement?

Referees report on 2025-12-32495A

April 30, 2026

This review attempts to walk a somewhat fine line. On the one hand, we continue to believe that the experimental work reported here is both high-quality and of significant and broad enough interest to merit publication in Nature. This work marks several significant steps forward for atomic clocks, at a time when the redefinition of the second amplifies the importance of atomic clock progress. We discuss this further below.

On the other hand, this revision has added a substantial focus on the authors' opinions about the state of atomic clock research and the shortcomings of the SI redefinition process. We have significant issues with these opinions and their inclusion in this report. The recommended changes conflate the clock evaluation with the more rigorous process of validation; they would likely not resolve the issues with comparison repeatability cited by the authors; the underlying thesis is applied in a logically inconsistent way; and the proposed criteria are not fully met by the experimental work presented here. We will also discuss each of these points in more detail below.

In sum, we do not believe that the manuscript should be published in its present form, but we hope the authors will revise and continue to seek publication in Nature. We believe that a version of this article which emphasizes the impact of the experimental work will satisfy other referees' concerns about broad interest, without centering the authors' opinions at the expense of their remarkable experimental achievement.

Other points are included at the end of the review, including concerns about the unmeasured secular-motion heating rate, where we note that published works and preprints have reported up to a factor of 5 anisotropy in macroscopic, symmetric traps.

Regarding impact and interest

Both of the other referees expressed the opinion that this result is not of sufficiently broad interest to serve Nature's readership. We disagree for several reasons, which we will summarize here.

First, there is the matter of multiple-record-setting performance. This is the most accurate clock ever constructed, by a factor of more than four. In addition to that evaluation, this work also reports the most precise clock comparison to date (and one of the most precise measurements of anything ever). We believe these records alone carry a certain amount of interest to physical scientists, and papers reporting record advancements in clock performance have been reported in Nature - e.g. Bloom et al., Nature 506, pages 71–75 (2014) reporting the JILA strontium lattice clock.

Second, this is the first comprehensive accuracy evaluation of a lutetium clock. This is an important milestone, since lutetium promises significant advances in accuracy compared to better-established clocks. In this sense, we find it analogous to initial evaluations of other types of clock previously published in Nature, e.g. King et al., Nature 611, pages 43–47 (2022) reporting the first comprehensive evaluation of a highly-charged-ion clock, as well as the many Nature reports of incremental progress towards a thorium nuclear clock. We note that this is especially important at a time when the CCTF is establishing the short-list of atomic clock candidates for the upcoming redefinition of the second; this work makes a strong argument for lutetium's relevance to that process. Perhaps this could be emphasized in a future revision.

Third, the combination of a systematic uncertainty evaluation with a same-species comparison is a significant advancement beyond the single-clock evaluations that the other referees compare to this work. While we disagree (see below) with the authors that same-species comparisons should be required before any clock evaluation can be published, it is indisputable that a same-species comparison makes a clock evaluation stronger; it bounds many types of possible experimental error, should make interspecies comparisons involving the clocks more reliable, and demonstrates clock performance in a more realistic scenario than an evaluation alone. We note that the only analogous work of which we are aware was also published in Nature, as McGrew et al, Nature 564, pages 87–90 (2018) reporting on the NIST ytterbium lattice clocks.

We also hope that the reorganization of the paper, with many technical details moved to the Methods section or supplemental material, and with important references and explanations added to many important points, will make it more accessible to non-specialists.

Issues with the revised manuscript

Having made the case that the scientific merits of this work rise to the high level expected from a Nature publication, we move to reasons we do not believe it should be published in its current form. We continue to have issues with the authors' discussion of "validation", as well as their new criteria for "acceptable" atomic clock evaluations. We do not find the authors have adequately addressed our concerns on this front; instead they have devoted significant new emphasis to their interpretation of "validation", which we find flawed, as well as to their criticism of atomic clock research more broadly.

Before getting into the specifics of this debate, we note that a peer-reviewed research report is not the appropriate forum for such opinion. This would conflate peer review of the research with peer review of the opinion. Large parts of the new "discussion" section are essentially unrelated to the research work being reported here, instead being part of the authors' criticism of the state of the field. Rather than inside a high-profile research report, we believe the community and the authors would be better served if these criticisms were published separately in a journal like Metrologia. That would be a more fitting venue for public debate; and the criteria could be presented as an aspirational standard for clock groups including the authors, rather than undermined by being published alongside a comparison which does not fully meet them.

There are a few separate issues here which we will break down.

Regarding "validation" as a term of art

The authors' new definition of "validation" is significantly less rigorous than the community consensus included in the SI redefinition roadmap, mandatory criterion I.2 (N Dimarcq et al., Metrologia 61 012001 (2024), ref. 2 of the revised manuscript). The core element lacking from the authors' definition is reproduction by independent experimental groups. "Validating" an atomic clock requires reproduction by different institutes, which is an essential check against experimenter error, subconscious bias, challenges from long-distance time transfer or clock transportation and unidentified shifts. This is fundamentally not something that can or should be done by a single institute.

For example, the authors cite the NIST Yb optical lattice clock (ref. 15 of the revised manuscript) as the only other atomic clock which meets their definition of a "validated" accuracy budget. We point them to the appendices of arXiv:2512.21428 (ref. 8 of the revised manuscript). Here, this same group discusses how, because of errors in atomic structure calculations, the systematic shifts applied in ref. 15 were too small by 3E-18 – more than double the total systematic uncertainty evaluated in ref. 15. Clearly, the presence of a same-species comparison with high precision did not guarantee that the accuracy budget was correct.

Additionally, the degree of independence of the two clocks matters when determining whether the evaluation can be considered "validated". As an example of the kind of unidentified shifts which are not eliminated in this comparison, we note that spectral impurities in a clock laser can cause – and have caused, in optical lattice clocks at least – significant systematic shifts, which would cancel in common mode in a comparison like the one performed here. Requiring that comparisons be between different institutions establishes an objective baseline for what can be considered "independent", which otherwise is highly subjective (as one can clearly see from our review and the authors' response).

This contention over "validation" is in part a semantic issue, since – as we discussed in our previous report – "validation" already has a specific definition in the context of next-generation clock uncertainty estimates. The authors have a legitimate claim that their uncertainty evaluation is better supported by evidence than others which do not include self-comparisons. A different, less loaded term would avoid this conflation of the two definitions – perhaps "confirmation", "verification", or "corroboration". However, both because of the common institution and because of the reduced level of independence of the two clocks, we continue to hold that this comparison doesn't reach the level of a fully "validated" frequency reference.

Does the lutetium clock meet the authors' criteria?

Semantics aside, we believe there are issues with the application of these criteria to the present work.

First, as applied to lutetium the criteria are logically inconsistent. By insisting that clock accuracy budgets are only "acceptable" insofar as they have been tested with same-species comparisons, the authors imply that rigorous evaluation via secondary measurements is insufficient or "unacceptable" – even where no specific reason exists to doubt such an evaluation. The argument is that to be trustworthy, these evaluations require additional information from a same-species comparison to corroborate the standard uncertainty evaluation. At core, this is an epistemological stance advocating skepticism towards a specific category of scientific results. For this stance to be meaningful, it must be applied uniformly. If only applied where convenient to the authors, it ceases to be rigorous.

The same-species comparison reported here provides no additional information to corroborate the evaluations of the blackbody shift or second-order Zeeman shift, since those are both common-mode between the clocks. Therefore, those aspects of the clock evaluation don't satisfy this epistemological criterion and should not be considered "acceptable".

In response to this point from our initial review, the authors claim that "both of these properties have been rigorously assessed". This is true; the scientific work in this manuscript is of high quality. However, systematic effect evaluations by many other high-quality research groups around the world also feature "rigorously assessed" properties with equally high quality. Insofar as the authors wish to discount those other accuracy evaluations because they haven't been tested in same-species comparisons, the exact same logic must apply to these shifts on lutetium. To be "acceptable", they must be corroborated by additional information from a same-species comparison. Applying this standard to an overall clock evaluation but doing so in a way that does not corroborate the two largest shifts is logically inconsistent.

We note that a same-species comparison at significantly different temperatures could, in principle, uncover errors in calculated or measured atomic properties; in that case, the incorrect BBR shift corrections applied for the two different temperatures would result in a non-unity ratio.

In this light, no clock evaluation can be "rigorous" under the authors' criteria until a same-species comparison has been made between two apparatus at significantly different temperatures and with significantly different magnetic fields. That this would be difficult is not sufficient reason to make an exception just for lutetium or just for the BBR shift; it's evidence that "validating" a clock without relying on rigorous characterization of effects through secondary measurements is, in fact, difficult.

We emphasize that we don't consider this a reason to doubt the evaluation of the BBR shift in this work. Rather, this undermines the authors' asserted criteria for clock evaluations and their criticism of standard practices in the clock field.

Second, the role of the same-species comparison in making an accuracy evaluation more rigorous is to constrain shifts due to unknown effects and due to experimenter error. A self-comparison can constrain these only insofar as they are not common-mode between the two clocks. However, the independence of these two systems is limited in ways that significantly reduce their ability to constrain unknown effects, even beyond the temperature and magnetic field servos. Shared environments, control software, data analysis pipelines, apparatus design, and other operating conditions are all avenues that unknown effects could come in through. To our knowledge, the majority of errors found in previous uncertainty evaluations would not have been caught by a same-species comparison similar to the one performed here; for example, the incorrect atomic structure coefficients or clock laser spectral impurities discussed above would both cancel in common mode.

One of the reasons the SI redefinition roadmap demands different institutions make comparisons is to control for these unknown effects. So-called "dark" uncertainty is unfortunately common in high-precision interspecies clock comparisons, both within comparison campaigns and between measurements of the same ratios (cf. references in section 6C of Fortier et al., Optica 13 143-163 (2026), as well as Aeppli et al., arXiv:2512.21428 (2026)). Same-species comparisons involving independent systems can help resolve these issues precisely by constraining unknown shifts affecting one or more clocks.

Third, this work doesn't meet the authors' own "minimum criteria for a rigorous assessment". For one, they are making a "performance claim" at the 1e-19 level but only do a comparison at the 5e-19 level. In the authors' language, the "claimed accuracy cannot be better than the measurement precision" for the same-species comparison. If the authors were reviewing a different group's clock accuracy evaluation, one assumes that they would implement this standard by refusing to accept an accuracy claim with precision beyond the relevant self-comparison. The same should be true here - if this performance claim is "unacceptable", it should be moderated to match this standard or the standard should be rethought.

For another case, the new criteria require "a stress test on any systematic that is more than an order of magnitude larger than the accuracy claim". As far as we can tell, the authors have not done sufficient "stress tests" on the two clocks presented here. Again, this would require changing the BBR environment significantly enough to evaluate the BBR shift. While the authors argue the differential BBR shift is small for a reasonable range of laboratory temperatures, what needs to be epistemologically confirmed and "stress tested" isn't the differential shift between the two clocks; it's the total BBR shift from 0K to room temperature, which is more than an order of magnitude larger than the claimed accuracy.

The measurements using different fields on Lu-1 (presented in Fig. 3) would satisfy this requirement for Lu-1 on the Zeeman shift, although it's not clear whether that applies to Lu-2 as well – that depends on whether the authors intend this to be a requirement only on a species-by-species basis, or separately for each clock evaluated.

Are these new criteria for "acceptable" clock comparisons a good idea in the first place?

The authors would like to add same-species comparisons to mandatory condition I.1, when they are already present in a more restrictive form in I.2. Aside from this being insufficient to demonstrate "validation" as defined in criterion I.2, we disagree that a same-species comparison should be required in order to report clock accuracy evaluations. We believe this stems from a different understanding of what an accuracy evaluation is or should be. As interpreted in nearly every other published optical clock result to date, the accuracy evaluation is nothing more nor less than a collection of measurements which rigorously estimate shifts and uncertainties to an atomic clock frequency. In other words, the authors' criterion i.

In essence: an accuracy evaluation is not a "hypothesis", as the authors put it in the revised introduction. It is a specific type of reported research result.

An accuracy budget does not amount to a guarantee of the variation in frequency ratio measurements from a "true" value – that includes the significant additional experimental effort which goes into making a high-precision comparison, including frequency transfer, phase-noise compensation, operation of optical frequency combs for inter-species comparisons, relative geopotential measurements between clocks, and data analysis. All of these are opportunities for error which cannot be appropriately included in any single clock's accuracy evaluation.

Adopting the authors' criteria would amount to invalidation of every previously published accuracy evaluation except one, which features identified flaws. Additionally, it would push institutions to redirect resources towards replicating identical apparatus rather than developing new clocks or advancing clock performance, and it would exclude groups without the resources to build two identical clocks.

Moreover, any additional utility from this definition is already captured by the existence of mandatory criterion I.2 – as part of the SI redefinition process, clock error budgets must be validated at 5E-18 by multi-institute clock comparisons. Beyond the redefinition, in the case of the optical second, the CCTF Working Group on Primary and Secondary Frequency Standards would continue to stringently evaluate individual frequency standards at the relevant level of precision, just as they do now for both microwave and optical standards at the E-16 level relevant to TAI.

We agree with the authors that same-species comparisons have additional value beyond the accuracy evaluation. They represent a check against many - but not all - potential significant errors in clock construction and evaluation. They also demonstrate the performance of a clock in a more direct way. Hence, again, why we support publication of the underlying results here in Nature rather than a less prestigious subject-matter-specific journal. However, this utility is additional to the accuracy evaluation, not inherent to it.

Magnetic field correlation analysis

We agree that the magnitude of the magnetic field noise will not be a significant issue, and the added details make this claim more convincing. Our comments and questions have been addressed. However, the newly added Fig. S2d requires some additional details. First, there is no metric provided for judging the magnetic field noise as 'uncorrelated.' One could claim a small anti-correlation based on visual inspection alone — what is the correlation coefficient? What does the thick black line represent? Second, the quantum projection noise, which is independent between Lu-1 and Lu-2, can dominate the correlation of the magnetic field noise. For the experimental conditions in Fig. S3a, the averaging time would need to be around 100 seconds to remove the projection noise component.

SODS and heating rate measurements

We were disappointed to learn that the authors had not measured the secular motional energy in one of the three secular modes for both clocks. The authors are correct about the distance scaling of anomalous heating, and that heating rates typically scale smoothly with frequency, but both significantly anisotropic heating and anomalous frequency scaling have been reported in other macroscopic traps. For example, Dawel et al (arXiv:2509.22525, 2025) observed a factor of 5 larger heating rate along one radial direction than the other, using a blade trap generally similar to that used in this work and with a difference in secular frequencies of less than 0.5% between modes. Lindvall et al (PR Applied 24, 044082, 2025) observed heating rates that changed by a factor of 5 over a measurement campaign, and depended on trap secular frequency beyond a smooth scaling. Leopold et al (Rev. Sci. Instrum. 90, 073201, 2019) observed a factor of 2 difference in heating rates for a cryogenic blade trap with a $\sim$10% difference in mode frequencies.

We note that while the trap used in this work demonstrates mechanical symmetry, there is not a matching electrical symmetry - according to ref. 50 of the revised manuscript, one pair of rod electrodes provides the RF pseudopotential and the other pair is held at DC. This implies different impedances to ground and different electrical sources connected to the pairs, which is a natural way anisotropic technical noise could affect the ion.

This is an area where the same-species comparison is valuable - realistic sources of anisotropic heating are likely technical in nature rather than inherent to the trap design, so they're unlikely to be correlated between the two systems. However, this only constrains the effect at the 5.7E-19 level, rather than the 1E-19 overall accuracy. Perhaps the authors can derive a more stringent constraint from the dephasing measurements, or measure the heating rate of the inaccessible mode using parametric mode coupling as demonstrated in Gorman et al (Phys. Rev. A 89, 062332, 2014). However, it's not obvious to us, given these results from a brief literature search, that potential anisotropy can be accounted for simply by increasing the uncertainty on the unmeasured mode from $\sim$5% to $\sim$10%.

First-order Doppler shift

We agree with the authors' discussion of charging and discharging from UV photoelectrons. This is an unrealistic possibility when applied to the Lutetium case. The point of bringing up the NIST group's measurement was not to suggest a particular source of the FODS, but rather to make the point that large first-order Doppler shifts can arise in ion trap systems and be difficult to detect without counterpropagating beams. The authors point out an issue with the photoelectron hypothesis from that experiment - if they are correct, that would only further emphasize that first-order Doppler shifts can arise without an obvious mechanism, and should be constrained by measurement.

Photoelectrons aside, the authors have three responses to our concerns about the potential first-order Doppler shift, which we will summarize here for clarity of discussion.

1. The self-comparison constrains any FODS, because the odds that it would be identical and synchronized between clocks are vanishingly small
2. The authors dispute the likelihood of the specific possible FODS scenarios we discussed in our initial report.
3. Constraining the FODS with new measurements using counterpropagating beams would take too long to average down, and "incorporating a purely hypothetical FODS into the uncertainty budget lacks scientific utility"

We will address each of these points in turn. Regarding the first point - we agree that the odds of an undetected FODS are small, and it is reasonable to constrain it at the level of the self-comparison. However, this comparison is five times less precise than the claimed systematic uncertainty, so this point is only partially relevant. This is related to the contradiction discussed above between the authors' performance claims and their epistemological stance on the necessity of self-comparisons.

Regarding the second point - we initially note that all of these proposed mechanisms were intended to illustrate the difficulty of constraining an FODS from first principles, rather than detailed estimates. In line with the authors' epistemological arguments, first-principles arguments such as these are relatively unsatisfying. A few specific notes:

- We agree that scattering from the ion is negligible. To our knowledge, the MW power used is not given in the manuscript; at single-digit milliwatts, we agree that is also unlikely to cause substantial differential heating across the trap.
- The optical powers used for interrogation were similarly not given in the manuscript. If the figure of 10 μJ per clock cycle includes laser cooling and readout as well as clock interrogation, then we agree the total power is likely too small to cause a substantial FODS. We appreciate the authors' provision of additional details.
- Regarding Lorentz forces, the concern is not oscillating currents, which as the authors point out average to zero for currents at different frequencies, but rather DC forces from power supply currents and ground return

currents, which may create a different electric-field environment when the laser or microwaves are off vs on. This is also likely too small to matter but again requires microscopic knowledge of the experimental apparatus, e.g. where do ground currents flow / how well have ground loop currents through the vacuum chamber been eliminated, etc.

- The timescale argument is only partially convincing. For example, suppose there is a mechanical vibration on the milliseconds scale from a shutter closing. To average 30 pm/s over the seconds-long probe time, the scale of the vibration at the ion or at one of the clock laser beam-path mirrors would have to be 30 nm/s, which still does not seem unreasonable if the shutter is close by and depending on the spring tension in the relevant mirror mounts.

- Relatedly, we note that ion clock experiments have observed first-order Doppler shifts related to the transient response of path-length servos to mechanical-shutter-induced vibrations. We believe this is already constrained by the AOM phase chirp measurement, but mention it to point out that the timescale question can be quite complicated - in this case, the transient response extends the timescale of the perturbation significantly beyond the timescale of the mechanical vibration.

Regarding the authors' third point - we fail to see a meaningful difference between incorporating a "purely hypothetical" error source into an accuracy budget just because it isn't experimentally constrained, and rejecting rigorously evaluated uncertainty budgets that don't come with a self-comparison to reject other purely hypothetical effects.

That aside - while it may be unsatisfying, we acknowledge that the standard practice in the field is to omit the FODS from uncertainty budgets unless there is specific reason to include it. We believe this leaves unanswered the question of whether large FODSes of unknown origin, similar to that observed by the NIST group referenced earlier, may also be present on other ion clocks. We believe this work would be better-supported had the authors conducted their campaign with counterpropagating clock beams. However, we acknowledge that they are reasonable at the level of standard practice in the field, especially when combined with the additional information provided in their response but not included in the manuscript, and don't require any additional changes to the manuscript on this front. We appreciate the inclusion of bounds on some FODS sources that can be estimated from the data the authors did collect.

That said, there is a certain tension with the authors' arguments about what constitutes an "acceptable" or "unacceptable" accuracy claim, where they discount rigorous evaluation and first-principles arguments as unverified without a self-comparison. Are the authors truly satisfied with this evaluation of the first-order Doppler shift, and if so what differentiates it from other clocks' "rigorous evaluation" of systematic effects which the authors consider "unacceptable?" The measurement of the first-order Doppler shift using counterpropagating clock beams would constrain otherwise un-measureable effects in the same way that a self-comparison would constrain some unknown differential effects; however, the authors decided not to make this measurement, while simultaneously claiming that accuracy budgets are not acceptable with uncertainty below the stability of a self-comparison.

If the authors want to present their clock as not just a significant experimental achievement, but as a model for scientific rigor and example of a framework that other atomic clocks should follow, they must also address the places where their work is in tension with that framework. Similarly to the lack of "stress tests" discussed earlier in our review, the fact that many potential sources of FODS are constrained only by first-principles argument rather than measurement should be addressed, or else the framework and epistemological arguments should be modified to be consistent with the experiment.

Small comments

Minor typo on page 2 of supplemental material: "We take this power to be half the total rf drive power assuming equal dissipation at each end of the trap **and**, i.e. all the power is dissipated in the ceramic."

Regarding the BBR shift: We appreciate the additional detail included in the new supplemental material. The authors have adequately supported their estimates of the BBR shift.

Similarly, we find that the questions about the magnetic field servo have been addressed.

Reviewer #4 (follow-up, July 16, 2026) for manuscript entitled, "Optical clocks with accuracy validated at the 19th digit" by K.J. Arnold et al

I appreciate the substantial effort the authors have made in revising the manuscript. The presentation has improved in several respects, and several of the more speculative claims have been moderated. However, I remain unconvinced that the manuscript meets the threshold for publication in *Nature*.

My principal concerns are as follows:

Accessibility to a broad scientific readership: Despite the revisions, I do not believe the manuscript is sufficiently accessible for *Nature*. The paper remains highly specialized and assumes considerable prior knowledge of optical clock operation, systematic uncertainty evaluation, and frequency ratio measurements. Much of the motivation, experimental work, and results and impact are presented at a level of specialization that I believe will be difficult to follow even for researchers in neighboring fields such as photonics or laser science. Given *Nature*'s broad readership, substantially greater effort is required to explain why this experiment matters beyond the optical clock community and what broader scientific question is being addressed.

Significance of the reported performance: The manuscript reports a comparison between two Lu+ optical clocks with a systematic uncertainty approaching 1×10^{-19}. While this represents an impressive technical achievement, the reported measurement uncertainty of approximately 5.7×10^{-19} limits the extent to which this systematic performance can presently be validated experimentally.

It is therefore unclear that the advance over the existing state of the art is sufficiently compelling for *Nature*. As one benchmark, the 2018 *Nature* demonstration comparing two independent ytterbium optical lattice clocks reported a systematic uncertainty of 1.4×10^{-18}, a measurement instability of 3.2×10^{-19}, and reproducibility from blinded comparisons at the 10^{-18} level. While the present work substantially reduces the evaluated systematic uncertainty, the experimental comparison itself is performed with a measurement uncertainty of approximately 5.7×10^{-19}. It is therefore less clear that the work establishes a correspondingly new level of experimentally demonstrated clock agreement.

I would also like to acknowledge an important distinction between the present work and the 2018 ytterbium lattice clock comparison. The Lu+ system is based on a single trapped ion, making this the highest-accuracy comparison reported for a single-ion optical clock to date. Compared with neutral-atom optical lattice clocks, which interrogate approximately 10^4 atoms simultaneously, single-ion systems inherently exhibit lower signal-to-noise ratios and higher quantum projection noise, resulting in reduced measurement stability. From this perspective, the work represents a significant technical milestone for single-ion optical clocks. My comments above concern whether the level of experimentally demonstrated advance, when viewed in the broader context of the field, reaches the editorial threshold expected for publication in Nature.

Personally, I think the most compelling demonstrations at this level of uncertainty are likely to be comparisons between fully independent clock systems (interspecies or same species) operating below the 10^{-18}level. Such experiments provide a more stringent validation of clock accuracy because they minimize the possibility that shared local oscillators, laboratory environments, experimental infrastructure, or correlated systematic effects contribute common-mode biases. Although this is not a requirement for publication, it illustrates why I find the overall significance of the present advance somewhat less compelling than the manuscript suggests.

In summary, the work represents a significant technical accomplishment and will likely be of considerable interest to the precision metrology community. However, in my opinion, the combination of limited accessibility for a broad readership and the degree of experimentally demonstrated advance relative to previous landmark optical-clock comparisons means that the manuscript does not currently meet the editorial threshold expected for publication in *Nature*.