

SuperARC: A Test for Artificial Superintelligence Based on Compressed Modelling, Recursive Prediction and Problem Complexity

Corresponding Author: Dr Hector Zenil

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The paper proposes a way to test a specific aspect of intelligence of artificial computer systems. The authors argue that the ability to compress strings should be included in the testable scope of intelligent behaviour, as it accounts for comprehension and abstraction abilities.

The authors then move on to comparing the existing LLM-based technologies on compression tasks. They juxtapose their performance with the technology combining Coding Theorem Method and Block Decomposition Method (which they call ASI - Artificial Super Intelligence). Indeed, as it turns out, the performance of CTM/BDM is far superior on the specific compression tasks.

The description of the techniques used and the methodology of the experiment is very well described. The experiment is reproducible, modulo the ever-increasing abilities of LLM and their dependence on particular data sets they are trained on. The work is original and well-motivated, and it builds upon previous work of the authors. The technical and mathematical standard of the work of high and the paper and the technical supplement are very readable.

The major weakness of the paper, as I see it, is in the interpretation of the results. In particular, the comparison between CTM/BDM and the existing LLMs seems unfair. LLMs might not perform well on these tasks, but they are useable across a variety of other problems on which the performance of CMT and BDM can't be sensibly tested. Another issue is that the authors link their results to real human intelligence. The usual question appears, what are we aspiring to? I am sure humans will not be perfect at data compression either.

Below some more specific comments.

Sentence in lines 40-43. Is it really the case that LLMs can actually show anything about human intelligence? This sentence seems to suggest that.

Line 63: what are "intelligent datasets"?

Sentence starting in line 75, unclear grammar.

Sentence starting in line 86, citation needed, the claim should be substantiated.

In Experimental design, lines 181-185: Please elaborate on what it means in practice that "this choice is not mandatory". Related to that, how would a different choice lead to "biasing LLMs"? Would that bias be to their advantage? If so, is selecting the most difficult set-up discriminates against the LLM technologies?

Line 234, instead of "Non-Prints & Non-Ordinal" could we say "Non-both", as the phrase was introduced previously.

Line 237, it is said ρ will use percentages, but later it's proportions, i.e., 1 instead of 100% (minor comment).

Lines 274-278, what is the model's "true performance" for LLM if data is not taken into account? It seems that overall criticism needs to be calibrated better. Arguably, all abilities of LLMs come from linguistic data.

Lines 283-288, the passage relies too much on the supplementary information, it should be rewritten to be more self-contained.

Lines 295-303, please elaborate on the possible source of the dominating performance of Llama.

There seems to be something wrong with the typesetting of the right quotation mark: line 188, 217, 229.

Please proofread bibliography, I noticed some typos there.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

This paper introduces SuperARC, a novel benchmark grounded in algorithmic information theory to assess the capabilities of large language models (LLMs). The key idea is to use the ability to compress and subsequently predict elements of a string as a meaningful proxy for intelligence. The idea is inspired from Kolmogorov complexity, which defines complexity of an object as the length of the shortest code that produces the object as output.

I find the core idea compelling: that compression ability may serve as a necessary (though perhaps not sufficient) condition for a form of genuine understanding. This shifts the evaluation focus from human-centric benchmarks to more universal tests. However, I remain uncertain about the practical implications of this benchmark for current LLM development. The paper would benefit from clearer articulation of how SuperARC scores might guide future architecture design, training methodology, or evaluation pipelines. In what ways might a model's performance under this framework concretely inform the development of more capable or general systems?

I agree with the critique of human-centric benchmarks, particularly the concern around benchmark contamination. However, a more direct empirical comparison with existing benchmarks would strengthen the argument. While SuperARC is designed to be contamination-resistant, it is not the only possible approach. For instance, benchmarks based on continuously updated or domain-specific datasets (e.g., newly submitted papers or reviews) may also sidestep contamination. It would be helpful for the authors to clarify what unique insights SuperARC provides that are not captured by such alternatives.

Another concern is the interpretation of performance. If efficient heuristics—not designed to reflect intelligence—can perform well on compression tasks, does that undermine the link between compression and understanding? Similarly, methods such as BDM and CTM are computationally intensive and not optimized for efficiency. Does their superior performance reflect deeper understanding, or merely exhaustive search? The paper could benefit from further conceptual clarity on what constitutes "genuine understanding," and whether that concept is equivalent to—or merely correlated with—compression success.

In summary, this work makes a thought-provoking contribution to the discourse on intelligence evaluation in artificial systems. However, to maximize its impact, the authors might consider clarifying the practical applications of SuperARC, offering comparative insights with existing evaluations, and elaborating on the philosophical implications of equating compression with comprehension.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for addressing my concerns! I am convinced that after the slight rephrasing, your work will constitute an impactful contribution.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I appreciate the authors' rebuttal and the revisions. I still find the central idea—evaluating AI models through abstract sequence prediction tasks conceptually interesting and potentially extensible to a broad family of tasks. I also agree that the ability to generalize across diverse domains of time series is important.

Though I became more positive after reading the revision. I am still not sure that the proposed approach, while elegant in its unification of core principles, represents a sufficient conceptual or empirical advance. From one perspective, human-centric tasks are already instantiations of this same predictive paradigm, as essentially those tasks are also prediction tasks. Moreover, the paper does not yet provide sufficiently strong evidence that this new benchmark would meaningfully shift current training or evaluation practices, nor does it demonstrate clear, policy-level implications.

In summary, while I appreciate the originality and the improved clarity in the revision and have become more positive toward the manuscript, I still hold the above reservations.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The paper introduces SuperARC, a benchmarking framework based on algorithmic information theory and algorithmic probability, intended to measure intelligence (whether human, AGI, or ASI) through a system's ability to compress data and use this compression for prediction. The authors test several recent large language models and show that, despite strong performance on human-language tasks, these models struggle with algorithmically complex sequences. This suggests that current AI systems rely more on pattern matching and memorisation than on deeper abstraction. This is an important question, because many public and scientific discussions assume that AI already performs genuine reasoning. Understanding where we really stand is essential for science, the economy, and society.

Overall, the paper is well written and presented. The core takeaway that today's LLMs fail to demonstrate strong algorithmic reasoning is important and well-supported by the benchmark and experimental results. The benchmark tasks are novel and interesting, and the failures are noteworthy and relevant to the broader debate about whether large models possess meaningful general intelligence. The idea of grounding the evaluation in algorithmic information theory rather than relying on conventional language benchmarks (designed for humans) is refreshing. Compression-based tasks have long been proposed as a way to measure structure discovery, and applying them in this context is useful and timely.

However, there are also downsides. First, the conceptual framing is somewhat narrow and employs (kind of) non-standard definitions of AGI, ASI, and intelligence. While the authors acknowledge this, redefining such terms risks confusing readers, especially when the examples given, such as calling a calculator a "narrow superintelligence," conflict with the usual meaning of "superintelligence" as broad and general (actionable-)cognitive superiority. The example in the supplement ("asking a human to behave like a cockroach") does not clearly support the definitions either, because humans can perform such behaviour at least at a principle level, whereas current AI systems cannot do so across arbitrary tasks. Moreover, the focused definition raises questions about who decides which tasks matter and how they relate to intelligence, a point that is unfortunately not addressed in the submission and that requires a diverse set of perspectives, not only a technological one. Actually, the paper reduces intelligence to prediction, which is too narrow. Biological and embodied cognition involves physical interaction, goals, adaptation, energy constraints, and ecological niches. Intelligence in plants, fungi, animals, and swarms arises from distributed or morphological processes that cannot be reduced to prediction alone. Thus, the paper adopts a simplified AI-theoretic view that is influential in some areas but not a consensus, and this should be stated clearly and justified.

Some claims in the introduction are also overstated or potentially rather anecdotal. The idea that stochastic gradient descent approximates a "Kolmogorov search" is not correct without very strong assumptions; even small datasets can break this interpretation, as we may also need background knowledge or other inductive biases. However, the "in the limit" is not discussed, nor are the implications of measuring intelligence. Similarly, the claim that LLMs are Turing-equivalent depends on extra external or novel mechanisms, e.g., for decoding that current models are likely not to use. As noted even in the cited Schuurmans et al. work, such equivalence requires generalised autoregressive decoding with unbounded sequence extension, which seems to go beyond how today's models operate, as well as the models evaluated in the submission. These statements should be qualified or removed.

Another major gap is that the paper does not explain what it would mean if a system passed SuperARC and when this is even the case (10 tasks? 100 tasks? 1000 task?) Given the limitations mentioned below of CTM and BDM, passing the benchmark may not imply general intelligence or ASI. The method is suitable for ruling out claims of deep algorithmic competence but cannot serve as a positive test of ASI in general. Without discussing this limitation, the paper risks giving the impression that intelligence reduces to compression, which is not justified when it comes to acting and social interactions (as the body as well as the societal constraints shape intelligence, too) and excludes many important aspects of real-world cognition. This is especially relevant because the paper spends many pages discussing AGI and ASI from a perspective that

overinterprets what the benchmark can measure.

Moreover, the authors state that SuperARC can use “any complexity-based metric,” but they only present one metric without comparing alternatives or explaining why it was chosen. Can one argue via equivalences of metrics? Nevertheless, CTM is uncomputable for all but the smallest objects, and BDM is an approximation with known weaknesses: sensitivity to block size, dependence on decomposition decisions, limited ability to capture long-range dependencies, and the mixing of statistical and algorithmic notions. These limitations should be discussed explicitly, because they directly affect the interpretation of the results, and maybe also experimentally investigated?

More broadly, the paper does not sufficiently address (potential) methodological pitfalls. It is unclear how sensitive the results are to prompting choices or alternative formats. It just presents one prompt (see page 24, “for which the LLM is presented with the following task” or is this not even a prompt?), at least in the main text. The claim that “any dataset can be used” overlooks the challenge of selecting appropriate data (sub)sets for testing algorithmic reasoning (coverage). The discussion of the information non-increase theorem also does not consider that a finite set of tasks may fail to detect algorithmic competence even if the model could, in theory, succeed on other algorithms. In other words, how do the selected tasks align with the competences required in the real world (as required for ASI), given that there is likely a real-world manifold? The paper also does not touch upon existing literature on LLM failures in program synthesis and algorithmic tasks, even though this work strongly relates to known limitations and, in turn, to SuperARC itself.

The paper is also far too long for Nature Communications. Sections like Section 3 read more like a textbook introduction to algorithmic information theory, covering material that has been discussed for decades by Solomonoff, Schmidhuber, Hutter, and others. Much of this background can be moved to supplementary material or shortened significantly. The manuscript does not follow the structure expected for Nature Communications, and a future review would require substantial re-reading and re-reviewing.

Some presentation issues also limit clarity. The description of SuperARC is very technical and should begin with a more intuitive explanation, with the detailed algorithm moved to an appendix. Table 1 is difficult to read and could be improved with clearer extra formatting (bold, bullets) or color. Furthermore, if SuperARC is intended to measure intelligence, even in the authors’ limited sense, it would be informative to discuss how humans perform on these tasks or whether a small user study could be meaningful (or not)

To summarize, the paper tackles an important question and provides interesting negative results showing that current LLMs lack deeper algorithmic reasoning capabilities. However, the framing, definitions, theoretical claims, methodological justification, and overall structure need substantial revision. The work would benefit greatly from being more to the point / shorter, clearer, more balanced, and more transparent about its assumptions and limitations. The submission is promising, but it requires a significant rewrite to make the argument easier to understand and appropriate for publication.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I think the paper is mature enough to be publish. Looking at the discussion, I agree that many points remain controversial, but I doubt that the paper can be significantly improved with respect to those.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I thank the authors for their detailed rebuttal and the substantial revisions made to the manuscript. It addresses some of my concerns. After reading the new discussions, I believe the benchmark is a timely and valuable contribution to the field of AI evaluation. In summary, the paper is most successful as a critique of current LLM drawbacks. It is less successful, in my view, as a definitive metric for AGI.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

This is a resubmission. The main issues raised in my previous review was summarized as “the framing, definitions, theoretical claims, methodological justification, and overall structure need substantial revision. The authors did a very good

job in addressing my concerns, making major changes/revisions to Sections 1, 6.2, 6.4, 7.2, and 7.3. Doing so really improved the paper. I very much appreciated this and I think the paper is ready for publication. While I would still question the underlying "school" of what intelligence is, this is a rather subjective matter and should not influence the decision.

(Remarks on code availability)

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/>

Manuscript title: SuperARC: Can Increasing Complexity Explain Intelligence? A Test for Artificial Super Intelligence Based On the Principles of Causal Recursive Compression and Algorithmic Probability

1 Response to Reviewer #1's comments:

1.1 Point:

The major weakness of the paper, as I see it, is in the interpretation of the results. In particular, the comparison between CTM/BDM and the existing LLMs seems unfair. LLMs might not perform well on these tasks, but they are useable across a variety of other problems on which the performance of CMT and BDM can't be sensibly tested.

Response:

In order to fully address the Reviewer's comments, we have revised the manuscript, and substantially modified specially Sections 2, 3, and 4. In these sections, new content was added and the ordering of the sections and structure of the text were reorganised so as to improve the streamlined readability while answering the Reviewer's comments.

In particular, we highlight the new content that was added into new Section 2 ("Assessing the capabilities of frontier models and LLMs"), Section 3.1 ("Compression and machine learning"), Section 3.3 ("Randomness, prediction, and compression"), Section 3.4 ("Compression as comprehension"), Section 4 ("SuperARC testing framework"), and Section 4.1 ("A neurosymbolic approach to a superintelligence benchmark"). In the marked pdfdiff, we have inserted commented orange balloons in these sections indicating where each topic was addressed.

In summary, we acknowledge that the reviewer's concern is valid. We have clarified that our analysis focuses on two specific features increasingly linked to intelligence, and that while LLMs are powerful linguistic tools, BDM and CTM are not intended to compete with them. We also emphasize that the other tests capture complementary aspects of intelligence and serve to enrich our approach when those additional features are of interest.

1.2 Point:

Another issue is that the authors link their results to real human intelligence. The usual question appears, what are we aspiring to? I am sure humans will not be perfect at data compression either.

Response:

We have substantially modified Sections 2, 3, and 4, to address this concern.

In particular, as indicated in the comments left in the pdfdiff, the Reviewer's comments were addressed mostly in the new Section 3 ("Algorithmic Information Theory (AIT) and Intelligence") and Section 4 ("SuperARC testing framework").

One of the main points we have aimed to make throughout this work is that compression is closely connected to comprehension through model selection and prediction. The prevailing view of intelligence—particularly human intelligence—frames it as a process of optimal planning. Our central thesis is that planning involves constructing computable models that act as simulators and predictors of future states, and are therefore fundamental to planning itself. In this sense, humans indeed strive for optimal, recursive (rather than merely statistical) compression. Conversely, even if such optimization is not an explicit human goal, it can be argued that it represents a central goal of science, and potentially a legitimate objective for AGI or ASI.

1.3 Point:

Sentence in lines 40-43. Is it really the case that LLMs can actually show anything about human intelligence? This sentence seems to suggest that.

Response:

We have revised and corrected the misleading text. In particular, to address the Reviewer's comment and clarify the terminology, this topic is fully addressed in the new Sections 2 and 3.

We replaced this term by an explanation of what we meant, which was simply a dataset of sequences or otherwise that have causal origins.

1.4 Point:

Line 63: what are "intelligent datasets"?

Response:

We have revised and corrected the text. We also added an excerpt in this part of the text to clarify the terminology, as recommended by the Reviewer. In the marked pdfdiff, we have inserted commented *orange* balloons in these sections indicating where this topic was addressed.

1.5 Point:

Sentence starting in line 75, unclear grammar.

Response:

This part of the text was revised and corrected.

1.6 Point:

Sentence starting in line 86, citation needed, the claim should be substantiated.

Response:

Citations have been added.

1.7 Point:

In Experimental design, lines 181-185: Please elaborate on what it means in practice that "this choice is not mandatory". Related to that, how would a different choice lead to "biasing LLMs"? Would that bias be to their advantage? If so, is selecting the most difficult set-up discriminates against the LLM technologies?

Response:

We have revised and corrected the text in order to disambiguate and clarify the terminology requested. In addition, to fully address the Reviewer's comments, these questions are addressed and indicated in the new Sections 3 and 4.

Because the test is designed to test AI and therefore also and especially LLMs on 'reasoning' capabilities understood in terms of model abstraction and prediction and not memorisation, we stressed a potential bias that was not designed on purpose but turned out useful to potentially identify the bias against memorisation, which indeed would be in detriment to LLMs as they rely heavily on memorisation. However, this bias was not by design, its identification was for good practice, and it only goes in the direction that we transparently want to test.

1.8 Point:

Line 234, instead of "Non-Prints & Non-Ordinal" could we say "Non-both", as the phrase was introduced previously.

Response:

We have revised and corrected this.

1.9 Point:

Line 237, it is said that ρ will use percentages, but later it is proportions, i.e., 1 instead of 100 (minor comment).

Response:

We have clarified the notation.

1.10 Point:

Lines 274-278, what is the model's "true performance" for LLM if data is not taken into account? It seems that overall criticism needs to be calibrated better. Arguably, all abilities of LLMs come from linguistic data.

Response:

We did not intend any pejorative connotation with the term ‘true’, so we have removed it to avoid any misunderstanding.

1.11 Point:

Lines 283-288, the passage relies too much on the supplementary information, it should be rewritten to be more self-contained.

Response:

We have made our best effort to make it more self-contained. On the other hand, we are handcuffed by strict length restrictions by the guidelines.

1.12 Point:

Lines 295-303, please elaborate on the possible source of the dominating performance of Llama.

Response:

We have added a comment, although we have no explanation for the better performance of Llama against the others.

1.13 Point:

There seems to be something wrong with the typesetting of the right quotation mark: line 188, 217, 229.

Response:

These and all punctuation marks throughout the manuscript were revised and corrected.

1.14 Point:

Please proofread bibliography, I noticed some typos there.

Response:

We have revised this section and the entire manuscript.

2 Response to Reviewer #2's comments:

2.1 Point:

I find the core idea compelling: that compression ability may serve as a necessary (though perhaps not sufficient) condition for a form of genuine understanding. This shifts the evaluation focus from human-centric benchmarks to more universal tests. However, I remain uncertain about the practical implications of this benchmark for current LLM development. The paper would benefit from clearer articulation of how SuperARC scores might guide future architecture design, training methodology, or evaluation pipelines. In what ways might a model's performance under this framework concretely inform the development of more capable or general systems?

⋮

Another concern is the interpretation of performance. If efficient heuristics—not designed to reflect intelligence—can perform well on compression tasks, does that undermine the link between compression and understanding? Similarly, methods such as BDM and CTM are computationally intensive and not optimized for efficiency. Does their superior performance reflect deeper understanding, or merely exhaustive search?

Response:

In order to fully address the Reviewer's comments, we have revised the manuscript and substantially modified especially Sections 2, 3, and 4. In these sections, new content was added, and the ordering of the sections and structure of the text was reorganised so as to improve the streamlined readability while addressing the Reviewer's comments.

In this regard, we highlight the new content added into new Section 3.3 ("Randomness, prediction, and compression"), Section 3.4 ("Compression as comprehension"), Section 4 ("SuperARC testing framework"), Section 4.1 ("A neurosymbolic approach to a superintelligence benchmark"), and Section 4.1.1 ("Applicability of CTM and BDM to abstraction and planning in machine learning"). For other concerns related to this point, see the response of *Point 2.3* below. In the marked pdfdiff, we have inserted commented *orange* balloons in these sections indicating where each topic was addressed.

2.2 Point:

I agree with the critique of human-centric benchmarks, particularly the concern around benchmark contamination. However, a more direct empirical comparison with existing benchmarks would strengthen the argument. While SuperARC is designed to be contamination-resistant, it is not the only possible approach. For instance, benchmarks based on continuously updated or domain-specific datasets (e.g., newly submitted papers or reviews) may also sidestep contamination. It would be helpful for the authors to clarify what unique insights SuperARC provides that are not captured by such alternatives.

Response:

Other benchmarks, including the ones from the ARC challenge, are unfortunately not compatible with each other, but they are also very different in nature without having to make a side-by-side comparison. For example, the ARC challenge’s input is illustrations of possible transformations to images based on the expected output of human beings. In this test, however, there was no human intervention in picking, designing, or expecting one answer or another. While there was a process of selection, this was mostly for complexity classification. We have added an explanation about it in the manuscript.

2.3 Point:

The paper could benefit from further conceptual clarity on what constitutes “genuine understanding,” and whether that concept is equivalent to—or merely correlated with—compression success.

⋮

In summary, this work makes a thought-provoking contribution to the discourse on intelligence evaluation in artificial systems. However, to maximize its impact, the authors might consider clarifying the practical applications of SuperARC, offering comparative insights with existing evaluations, and elaborating on the philosophical implications of equating compression with comprehension.

Response:

Thanks for the positive comments.

In order to address the reviewer’s comments, suggestions, and questions, we have substantially revised Sections 2, 3, and 4. In these sections, new content was added (as indicated in the commented *orange* balloons), and we improved the ordering and structure of the text.

In particular, the respective new content and modifications to address this point can be found throughout the entire new Section 3 (“*Algorithmic Information Theory (AIT) and Intelligence*”), and in Section 4 (“*SuperARC testing framework*”), Section 4.1 (“*A neurosymbolic approach to a superintelligence benchmark*”), and Section 4.1.1 (“*Applicability of CTM and BDM to abstraction and planning in machine learning*”). In the marked pdfdiff, we have inserted commented *orange* balloons in these sections indicating where each topic was addressed.

We thank the reviewers for their constructive feedback and we are grateful they are mostly supportive of this work.

With kind regards. On behalf of the authors,

Dr. Hector Zenil
Associate Professor
King’s College London
(hector.zenil@kcl.ac.uk)

Response Letter – Revision 3

February 10, 2026

Submission NCOMMS-25-48295A

1

Manuscript title: Can Complexity and Uncomputability Explain Intelligence? SuperARC: A Test for Artificial Super Intelligence Based on Recursive Compression

General Comments

All the modifications can be tracked down from the *Markup version* provided in the submission of the revised manuscript.

Notice that the entire paper was revised and new content (including new sections) was added across both the main paper and the Supplementary Information in order to address the Reviewers' comments.

In addition to the point-by-point answers (*see below*), we would like to briefly point out the structural changes applied in this revision as a result of the feedback received from the Reviewers in order to match the journal guidelines. In particular, we highlight the following:

- old Section 2 ('Assessing the capabilities of frontier models and LLMs') was moved to new Section 6.1 under new Section 6 ('Methods').
- old Section 3 ('Algorithmic Information Theory (AIT) and Intelligence') along with its first three subsections was moved to new Section 7.1 in the new version of the *Supplementary Information*.
- old Section 3.4 ('Compression as comprehension') was moved to new Section 6.2 ('Compression as comprehension about (and as part of) the world') under new Section 6 ('Methods').
- old Section 4 ('SuperARC testing framework') was moved to new Section 6.3 under the new Section 6 ('Methods').
- old Section 4.1 ('A neurosymbolic approach to a superintelligence benchmark') was moved to new Section 6.5 under the new Section 6 ('Methods').
- old Section 4.1.1 ('Applicability of CTM and BDM to abstraction and planning in machine learning') was moved to new Section 6.5.
- old Section 4.2 ('A method for measuring comprehension via algorithmic probability') was moved to new Section 6.6 under the new Section 6 ('Methods').
- old Section 4.3 ('Design of experiments') along with its subsections was moved to new Section 6.7 under the new Section 6 ('Methods').
- old Section 5 ('Results') along with its subsections was moved to new Section 2.
- old Section 6 ('SuperARC-seq') along with its subsection was moved to new Section 2.4.
- old Section 7 ('Conclusions') was moved to new Section 3 ('Discussion').
- old Sections 10.3 ('Equivalence between compression and prediction via Martingales') and 10.4 ('Levin's Distribution and the Algorithmic Probability of Integer Sequences') were moved to new Section 7.1.4 and 7.1.5, respectively, under new Section 7.1.

1 Response to Reviewer #1's comments:

Thank you for addressing my concerns! I am convinced that after the slight rephrasing, your work will constitute an impactful contribution.

Response:

We thank the reviewer for once again reviewing the paper and we are glad that the changes properly addressed the concerns of the previous review round.

2 Response to Reviewer #2’s comments:

2.1 Point:

I appreciate the authors’ rebuttal and the revisions. I still find the central idea—evaluating AI models through abstract sequence prediction tasks conceptually interesting and potentially extensible to a broad family of tasks. I also agree that the ability to generalize across diverse domains of time series is important.

Though I became more positive after reading the revision. I am still not sure that the proposed approach, while elegant in its unification of core principles, represents a sufficient conceptual or empirical advance. From one perspective, human-centric tasks are already instantiations of this same predictive paradigm, as essentially those tasks are also prediction tasks. Moreover, the paper does not yet provide sufficiently strong evidence that this new benchmark would meaningfully shift current training or evaluation practices, nor does it demonstrate clear, policy-level implications.

In summary, while I appreciate the originality and the improved clarity in the revision and have become more positive toward the manuscript, I still hold the above reservations.

Response:

We appreciate the acknowledgement of interest and acknowledgement of our improved revision. We have carefully addressed the remaining reservations:

In response, we have made substantial revisions to more clearly articulate: (1) the fundamental distinction between our approach and existing predictive paradigms, (2) concrete pathways for integration into current AI development practices, and (3) policy-level implications for AGI/ASI assessment and safety. We address each concern below and indicate the corresponding changes in the manuscript.

More specifically:

(1) Regarding the concern that “human-centric tasks are already instantiations of this same predictive paradigm”. We agree that prediction is fundamental to both approaches, there exists a critical qualitative distinction that we believe is now made more explicit in the revised manuscript. Human-centric benchmarks test mimicry of statistically reproducible patterns from human-generated content. This reduces most tests to statistical pattern-matching over the distribution of human knowledge and language (see Section 6.4). Our Theorem 1 (Section 7.1.4 of Sup. Inf.) formalises this distinction: we prove that predictive power through arbitrary formal theories is directly proportional to compression over the algorithmic space, not the statistical space. This is conceptually different from the maximisation of likelihood over human-generated data. Such distinction was empirically demonstrated, for example, by noticing that from ChatGPT-4.5 to ChatGPT-5, improved human benchmark scores were observed, but degraded SuperARC performance. Also, our neurosymbolic baseline achieves superior SuperARC performance despite having no exposure to human language or knowledge bases. These divergences demonstrates that optimizing for human-centric prediction does not guarantee, and may even hinder, algorithmic generalization.

In addition to predictive capabilities, SuperARC also tests the capability to achieve (algorithmic) discovery which requires: abductive reasoning to hypothesize generative processes; compression

over algorithmic space (not statistical space); and generalization beyond memorized training distributions.

As our results demonstrate, models can regress in fundamental algorithmic reasoning (SuperARC scores) while simultaneously improving in human-centric benchmarks. This would be impossible if both measured the same underlying capability.

(2) Regarding integration into "current training or evaluation practices": We have added a new section (Section 7.13 "Practical Applications and Integration into AI Development" in the Sup. Inf.) that outlines concrete pathways for SuperARC adoption. In summary, we indicate that SuperARC can be used to identify whether architectural improvements yield genuine algorithmic reasoning vs. statistical memorization. Also, our framework suggests prioritizing a few steps during model development and training, such as synthetic data with verifiable algorithmic structure; and curriculum learning from simple to complex algorithmic patterns and explicit symbolic reasoning modules. We also argue that SuperARC provides an orthogonal evaluation axis to allow model comparison beyond parameter count. We propose a "dual-benchmark" protocol where models are evaluated on both human-centric tasks (for practical utility) and SuperARC (for algorithmic generalization). Models showing divergence (improving on one while degrading on the other) should trigger investigation into training data quality and memorization vs. reasoning ratios (see also the previous response in the above paragraph (1)).

(3) Regarding "policy-level implications": We have substantially expanded our discussion of implications for AI governance and safety (Section 7.14 "Implications for AI Policy and Governance" in the Sup. Inf.). We argue that SuperARC addresses a critical challenge in AI governance by putting forward a mathematically sound and applicable method for distinguishing general intelligence from narrow systems optimized for certain tasks. As AI systems approach human-level performance on traditional benchmarks, policymakers need robust assessment frameworks for safety protocols, resource allocation, and international coordination. Unlike current benchmarks that increasingly measure memorization rather than reasoning capacity, SuperARC provides a human-agnostic, open-ended evaluation grounded in algorithmic information theory that cannot be gamed through data contamination or targeted optimization because it also accounts for potential interaction between tested and evaluator agents (see Sections 6.2 and 7.2).

Additionally, our framework offers practical applications for AI policy through capability-based governance triggers that focus on demonstrated abilities rather than proxy measures such as previously assigned parameters. SuperARC's neutrality, language independence, and verifiability make it particularly suitable for international coordination and standards. We propose a tiered approach that ranges from standard protocols for current LLMs (human-benchmark proficiency only) to maximum scrutiny for hypothetical AGI systems (superhuman performance on both human-centric and algorithmic dimensions), directly addressing the risks policymakers prioritize.

Our finding that LLMs are "tools optimised for the perception of mastery" rather than genuine reasoning has critical policy implications, as apparent capability may lead to misaligned expectations, inadequate safety evaluations, and resource misallocation toward benchmark performance rather than fundamental reasoning abilities. As our new Section 7.14 proposes, the evidence reported in our paper suggests policy support should prioritize hybrid neurosymbolic architectures, training paradigms that enhance algorithmic generalization, and funding mechanisms that require reporting both human-centric and algorithmic reasoning metrics to incentivize balanced progress.

3 Response to Reviewer #3's comments:

3.1 Point:

The paper introduces SuperARC, a benchmarking framework based on algorithmic information theory and algorithmic probability, intended to measure intelligence (whether human, AGI, or ASI) through a system's ability to compress data and use this compression for prediction. The authors test several recent large language models and show that, despite strong performance on human-language tasks, these models struggle with algorithmically complex sequences. This suggests that current AI systems rely more on pattern matching and memorisation than on deeper abstraction. This is an important question, because many public and scientific discussions assume that AI already performs genuine reasoning. Understanding where we really stand is essential for science, the economy, and society.

Overall, the paper is well written and presented. The core takeaway that today's LLMs fail to demonstrate strong algorithmic reasoning is important and well-supported by the benchmark and experimental results. The benchmark tasks are novel and interesting, and the failures are noteworthy and relevant to the broader debate about whether large models possess meaningful general intelligence. The idea of grounding the evaluation in algorithmic information theory rather than relying on conventional language benchmarks (designed for humans) is refreshing. Compression-based tasks have long been proposed as a way to measure structure discovery, and applying them in this context is useful and timely.

However, there are also downsides. First, the conceptual framing is somewhat narrow and employs (kind of) non-standard definitions of AGI, ASI, and intelligence. While the authors acknowledge this, redefining such terms risks confusing readers, especially when the examples given, such as calling a calculator a "narrow superintelligence," conflict with the usual meaning of "superintelligence" as broad and general (actionable-)cognitive superiority. The example in the supplement ("asking a human to behave like a cockroach") does not clearly support the definitions either, because humans can perform such behaviour at least at a principle level, whereas current AI systems cannot do so across arbitrary tasks. Moreover, the focused definition raises questions about who decides which tasks matter and how they relate to intelligence, a point that is unfortunately not addressed in the submission and that requires a diverse set of perspectives, not only a technological one. Actually, the paper reduces intelligence to prediction, which is too narrow. Biological and embodied cognition involves physical interaction, goals, adaptation, energy constraints, and ecological niches. Intelligence in plants, fungi, animals, and swarms arises from distributed or morphological processes that cannot be reduced to prediction alone. Thus, the paper adopts a simplified AI-theoretic view that is influential in some areas but not a consensus, and this should be stated clearly and justified.

⋮

Without discussing this limitation, the paper risks giving the impression that intelligence reduces to compression, which is not justified when it comes to acting and social interactions (as the body as well as the societal constraints shape intelligence, too) and excludes many important aspects of real-world cognition. This is especially relevant because the paper spends many pages discussing AGI and ASI from a perspective that overinterprets what the benchmark can measure.

Response:

We thank the reviewer for the constructive review and detailed feedback that has been invaluable in helping us substantially improve the manuscript.

We have undertaken a major revision addressing each of your concerns, including structural changes in the Section ordering (see the above *General Comments*), in particular to Sections 1, 6.2, 6.4, 7.2, and 7.3.

Below, we respond point-by-point to your critiques and indicate the corresponding changes in the revised manuscript.

In the new revised manuscript, we clarify that the scope of our framework and results extends beyond the predictive success of science and mathematics, also to include the synthesis of new formal theories and the creation of models, and the interplay between the AI algorithms being tested and the agents testing them. As explained in the revised version, instead of trying to capture every aspect of (or all the sufficient conditions for) intelligence, these are necessary (although not sufficient) aspects of intelligent systems that are fundamental to scientific and mathematical creativity, and we assume them to be crucial for any intelligence measure that aims at general capabilities.

Beyond prediction or compression, as detailed in Sections 6.2 and 7.2, SuperARC is an interactional approach in its theoretical underpinnings and empirical methods that combines perturbation analysis with the principles of AIT. Our framework accounts for interventions or interactions among the observer, the observed phenomena, and other external agents that may influence this interplay. As pointed out by the Reviewer regarding other types of cognition, disregarding the capability of interaction between many agents has been demonstrated to restrict the range of aspects of intelligence that can be measured, which has also been a limitation of previous universal AI formalizations based on AIT in the literature, such as the ones introduced by Hutter, Schmidhuber, and others.

We have made sure that our test is depicted and understood better in this revision as proposing what we think is measuring a necessary but not necessarily sufficient condition for general or superintelligence.

We have also included in the revised manuscript the important and required role of human-centric benchmarks for tasks other than compression and other types of intelligent behaviour, as recommended by the Reviewer's comments. As introduced and clarified in Section 7.13 ('Practical Applications and Integration into AI Development'), we propose that future implementations of our framework should be a dual-purpose protocol where models are evaluated on both human-centric tasks and for algorithmic generalisation.

Building on these new conceptual and mathematical clarifications, the revised Section 7.3 ('Challenges in defining AGI and ASI') acknowledges the problems with the definitions of AGI and ASI identified by the Reviewer, while detailing our approach to addressing this challenge relative to the usual understanding of those concepts. We believe that with our new formalizations introduced in the revised manuscript, the notion of AGI and ASI that we employ encompasses a broad range of possible interpretations of those terms, at the same time allowing for a discussion on common ground in mathematical and formal-theoretic terms.

We have also removed the calculator example in the Supplementary Information as "narrow superintelligence" and the cockroach example to avoid conflating task-specific optimization with superintelligence.

3.2 Point:

Some claims in the introduction are also overstated or potentially rather anecdotal. The idea that stochastic gradient descent approximates a “Kolmogorov search” is not correct without very strong assumptions; even small datasets can break this interpretation, as we may also need background knowledge or other inductive biases. However, the “in the limit” is not discussed, nor are the implications of measuring intelligence. Similarly, the claim that LLMs are Turing-equivalent depends on extra external or novel mechanisms, e.g., for decoding that current models are likely not to use. As noted even in the cited Schuurmans et al. work, such equivalence requires generalised autoregressive decoding with unbounded sequence extension, which seems to go beyond how today’s models operate, as well as the models evaluated in the submission. These statements should be qualified or removed.

Response:

We agree with the Reviewer’s comments so that some of the claims cited could be better contextualized and presented, avoiding confusing the readers.

Regarding the stochastic gradient descent approximation to a ‘Kolmogorov search’, the text has been rewritten in Section 1 (‘Introduction’) and we have added the following:

Based on these arguments connecting intelligence to recursive compression [6], some tests for machine, human, and non-human entities have been proposed in [16,17,5]. Section 6.2 presents a reflection on that property of intelligence to involve the identification of recursive patterns, planning from prediction, and the generation of concise explanations for observed complex phenomena. Recursive compression here means the ability to represent an observation in a condensed manner by taking advantage of aspects of the data’s regularities beyond statistical pattern matching. This is, by selecting and keeping as many as possible the features that make the explanation executable and predictive of the explanandum future states.

Despite the interesting theoretical arguments that could be drawn from these connections, one argument is that they would only be valid under idealised conditions (unbounded data access/storage, perfect optimisation, appropriate inductive biases), which are rarely met in practice. As seen in real-world problems, even simple datasets with specific distributions can lead to optimisation toward local minima that do not correspond to minimal algorithmic descriptions.

Regarding the Turing Equivalence statement, we have updated Section 6.1 under the new Section 6 (‘Methods’):

While, in principle, LLMs have shown to be theoretically capable of Turing-complete computation [52,53], this is achieved when they are augmented with external memory and appropriate decoding mechanisms [53]. In practice, the models we evaluate operate with standard autoregressive decoding and finite context windows, which do not constitute Turing-complete systems.

3.3 Point:

Another major gap is that the paper does not explain what it would mean if a system passed SuperARC and when this is even the case (10 tasks? 100 tasks? 1000 task?) Given the limitations

mentioned below of CTM and BDM, passing the benchmark may not imply general intelligence or ASI. The method is suitable for ruling out claims of deep algorithmic competence but cannot serve as a positive test of ASI in general.

Response:

We thank the reviewer for this critical observation, which has led us to clarify our framework’s scope and interpretation. We agree that our previous versions of the manuscript could give the impression that intelligence reduces to compression and that SuperARC could serve as a definitive test of AGI/ASI. As also clarified in the above responses, we have revised the manuscript extensively to address those. In the response of Point 3.4 below, we address the limitations of CTM and BDM.

In the newly added Section 7.13.4, we have included a more explicit discussion and clarification regarding what it means when SuperARC is passed. This is not a test to be passed or failed with a fixed threshold, but rather a continuous evolving metric to be used alongside other pillars of intelligence assessment. SuperARC performance must be understood comparatively rather than absolutely. As recommended by the Reviewer, we have added an extensive discussion in Section 7.13.4 explaining these concerns.

3.4 Point:

Moreover, the authors state that SuperARC can use “any complexity-based metric,” but they only present one metric without comparing alternatives or explaining why it was chosen. Can one argue via equivalences of metrics? Nevertheless, CTM is uncomputable for all but the smallest objects, and BDM is an approximation with known weaknesses: sensitivity to block size, dependence on decomposition decisions, limited ability to capture long-range dependencies, and the mixing of statistical and algorithmic notions. These limitations should be discussed explicitly because they directly affect the interpretation of the results and maybe also experimentally investigated?

Response:

We acknowledge that our statement about “any complexity-based metric” could be clearer, and we have rewritten this part of the text in the revised manuscript. We have revised and added explicit discussion on BDM’s known limitations in Section 6.5 (‘A neurosymbolic approach to a superintelligence benchmarking’) under the new Section 6 (‘Methods’), including the relationship of those with the results presented in the manuscript.

Regarding experimental investigation of these limitations through systematic sensitivity analyses, we respectfully note that such extensive empirical work would constitute a separate research contribution beyond the scope of this paper, yet we have added some future direction pointers.

Our primary contribution is demonstrating that current LLMs fail at algorithmic reasoning as measured according to a new framework that tackles the challenge of investigating the notions of AGI and ASI at a formal-theoretic level. We agree that investigating the comparative performance of alternative complexity approximations is important future work, but we believe—as also clarified in Section 6.5—it is not necessary to support our current core findings.

3.5 Point:

More broadly, the paper does not sufficiently address (potential) methodological pitfalls. It is unclear how sensitive the results are to prompting choices or alternative formats. It just presents one prompt (see page 24, "for which the LLM is presented with the following task" or is this not even a prompt?), at least in the main text. The claim that "any dataset can be used" overlooks the challenge of selecting appropriate data (sub)sets for testing algorithmic reasoning (coverage). The discussion of the information non-increase theorem also does not consider that a finite set of tasks may fail to detect algorithmic competence even if the model could, in theory, succeed on other algorithms. In other words, how do the selected tasks align with the competences required in the real world (as required for ASI), given that there is likely a real-world manifold? The paper also does not touch upon existing literature on LLM failures in program synthesis and algorithmic tasks, even though this work strongly relates to known limitations and, in turn, to SuperARC itself.

Response:

Regarding prompting sensitivity, we acknowledge that we presented only one prompt format in the main text for clarity. We tested several variations during development and found that while absolute performance varied, our core comparative findings (regression patterns across model versions and the gap between LLM and neurosymbolic performance) remained consistent. Since identical prompts were used for all versions within each model family, the observed regressions cannot be attributed to prompt engineering choices.

We agree that our claim that "any dataset can be used" was overstated and have revised it to acknowledge the practical challenges of appropriate dataset selection. The reviewer correctly notes that our finite test set may miss competences that would appear on other algorithmic classes. We now explicitly discuss these problems in the new Section 7.13 ('Practical Applications and Integration into AI Development').

Regarding alignment with "real-world manifold" competences, the newly introduced section 7.13 also discusses about this topic.

About making links to existing literature on LLMs' failures in related tasks, we have included the following paragraph at the end of Section 2.4.1:

These results are aligned with recent work exploring, for example, LLMs' logical reasoning failures [42] as well as GPT-4's limitations in deductive reasoning [43]. Other researchers have also reported degradation of mathematical capabilities [44] and planning limitations [45]. These works collectively document that LLMs struggle with systematic reasoning across multiple domains.

3.6 Point:

The paper is also far too long for Nature Communications. Sections like Section 3 read more like a textbook introduction to algorithmic information theory, covering material that has been discussed for decades by Solomonoff, Schmidhuber, Hutter, and others. Much of this background can be moved to supplementary material or shortened significantly. The manuscript does not follow the structure expected for Nature Communications, and a future review would require substantial re-reading and re-reviewing.

Some presentation issues also limit clarity. The description of SuperARC is very technical

and should begin with a more intuitive explanation, with the detailed algorithm moved to an appendix. Table 1 is difficult to read and could be improved with clearer extra formatting (bold, bullets) or color. Furthermore, if SuperARC is intended to measure intelligence, even in the authors' limited sense, it would be informative to discuss how humans perform on these tasks or whether a small user study could be meaningful (or not).

To summarize, the paper tackles an important question and provides interesting negative results showing that current LLMs lack deeper algorithmic reasoning capabilities. However, the framing, definitions, theoretical claims, methodological justification, and overall structure need substantial revision. The work would benefit greatly from being more to the point/shorter, clearer, more balanced, and more transparent about its assumptions and limitations. The submission is promising, but it requires a significant rewrite to make the argument easier to understand and appropriate for publication.

Response:

In the revised manuscript, we have dramatically reduced the size from ~ 18600 words to ~ 8700 (which excludes the new Section 6 'Methods'); and we have restructured the paper in order to match the section headings required by *Nature Communications*.

As recommended by the Reviewer, the sections regarding the theoretical background of the paper were moved to the Sup. Inf. In particular, the mentioned material that has been previously discussed in the literature is now addressed in Section 1 and the new Section 7.2 ('An Algorithmic Information Dynamics of AI algorithms, external processes, and evaluator agents').

We have also moved most of the explanatory technical content to the new Section 6 ('Methods') according to the journal guidelines.

Regarding the SuperARC description, after introducing the new subsections addressing scope and limitations, interpretation of performance metrics, and methodological considerations, we are confident that the overall readability and comprehensibility of the paper have been substantially enhanced. Most of the introductory concepts of our proposal are included in Section 1, and the more technical explanations are located in the new Section 6.

The discussion of the comparison between SuperARC and the way humans perform tasks is now included both at the beginning of the Results section, and in Sections 7.2, 7.3, and 7.13. We now cite an experiment performed several years ago with over 3000 human participants from a PLOS paper using the same methods (<https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1005408>)

As requested, Table 1 has been reformatted with improved visual presentation. We have introduced bold formatting to highlight top-performing models within each metric column. These modifications provide clearer visual hierarchy and facilitate rapid identification of performance patterns across models and metrics.

We believe the revised manuscript is easier to read and significantly stronger for having addressed these concerns while clarifying its assumptions and limitations.

We thank the reviewers for their constructive feedback and we are grateful that they were mostly supportive of this work.

On behalf of the authors,

Dr. Hector Zenil
Associate Professor
King's College London
(hector.zenil@kcl.ac.uk)