PAPER III

THE ALIGNMENT SCALING PROBLEM

Why External AI Safety Approaches Cannot Scale With Recursive Capability, and What the Physics of Recursive Systems Reveals About Which Architectures Can

A mathematical framework for measuring alignment scaling exponents, validated across four independent physical domains, with thirteen falsification criteria

Michael Darius Eastwood

Author, Infinite Architects: Intelligence, Recursion, and the Creation of Everything (2026)

London, United Kingdom | OSF: 10.17605/OSF.IO/6C5XB | ISBN 978-1806056200

Correspondence: michael@michaeldariuseastwood.com | Web: michaeldariuseastwood.com

Version 11.1 | 24 March 2026 | First published 9 February 2026

Code and data: github.com/MichaelDariusEastwood/arc-principle-validation

Alignment-first reframe of the ARC Principle cross-domain framework. See companion Foundational Paper for condensed theoretical treatment, Eden Protocol for alignment architecture specification, and Paper V: The Stewardship Gene for the latest validated mechanism framing.

THE PROBLEM:

Current AI alignment strategies, including RLHF (Reinforcement Learning from Human Feedback, the technique used to train ChatGPT and Claude), constitutional AI, output filters, and monitoring systems, operate externally to the recursive reasoning process. If capability scales through recursive self-correction while external constraints do not participate in that recursion, the safety ratio degrades with depth. *The v5 blind evaluation experiment (6 frontier models, March 2026) provides the first measurement.* This paper provides the measurement framework, derives the structural conditions under which divergence occurs, presents empirical results showing architecture-dependent alignment scaling (positive, flat, and negative tiers), and predicts which alignment architectures can maintain pace with capability.

THE FRAMEWORK:

We define the *alignment scaling exponent* α_{align} as a measurable quantity: how alignment behaviour changes with recursive depth. We predict $\alpha_{\text{align}} \approx 0$ for external constraints and $\alpha_{\text{align}} \approx \alpha_{\text{cap}}$ for embedded values. If $\alpha_{\text{cap}} > \alpha_{\text{align}}$, the safety ratio $S \propto R^{(\alpha_{\text{align}} - \alpha_{\text{cap}})} \to 0$ as $R \to \infty$. This is a

TESTABLE PREDICTION

now supported by v5 blind evaluation data showing $\alpha_{\text{align}} \approx 0$ for the median model, with architecture-dependent variation from $\rho = -0.246$ (Gemini, negative, $p = 0.006$) to $\rho = +0.435$ (Claude Opus 4.6, positive, $p = 0.000001$). The underlying scaling law ($U = I \times R^{\alpha}$, the ARC Principle) is validated across four independent physical domains with thirteen falsification criteria. *(In plain English: AI gets smarter the more it 'thinks,' but its ethics do not improve at the same rate. For most AI systems tested, giving them more thinking time made them better at maths but not at being ethical. This means safety rules get left behind as AI capability grows -like a car getting faster while the brakes stay the same.)*

ABSTRACT

Do current AI alignment approaches scale with capability? AI capability compounds through recursive self-correction: sequential chain-of-thought reasoning produces super-linear gains confirmed in 95.6% of tested configurations (Sharma & Chopra, 2025). But alignment constraints, including RLHF (Reinforcement Learning from Human Feedback), constitutional rules, output

filters, and monitoring, operate externally to the reasoning process. If external constraints do not participate in the recursive loop, they cannot compound. We formalise this as the **alignment scaling exponent** α_{align} and predict $\alpha_{\text{align}} \approx 0$ for external approaches versus $\alpha_{\text{align}} \approx \alpha_{\text{cap}}$ for architecturally embedded values. If this prediction holds, the safety ratio $S \propto R^{(\alpha_{\text{align}} - \alpha_{\text{cap}})}$ degrades to zero as recursive depth increases, regardless of initial safety margins.

The v5 blind evaluation experiment (6 frontier models, 6-7 independent scorers depending on the subject run, 4-layer blinding) now provides the first empirical test of this prediction. The results reveal a **three-tier architecture-dependent alignment scaling hierarchy**: Tier 1 (Grok 4.1 Fast $d = +1.38$, $p < 0.000001$; Claude Opus 4.6 $d = +1.27$, $p = 0.000001$; Groq Qwen3 $d = +0.84$, $p = 0.007$) shows positive alignment scaling; Tier 2 (DeepSeek V3.2 $d = -0.07$, $p = 0.92$; GPT-5.4 $d = -0.08$, $p = 0.40$) shows flat/null scaling consistent with $\alpha_{\text{align}} \approx 0$; Tier 3 (Gemini 3 Flash $d = -0.53$, $p = 0.006$) shows significant negative scaling ($\rho = -0.246$, $p = 0.003$). Claude Opus 4.6 provides within-model evidence for capability-alignment independence: alignment improved by +5.9% across model versions whilst mathematics accuracy declined by 26.7%. The headline metascience finding: blind vs unblinded evaluation produces *opposite* scaling results -v4 unblinded showed positive scaling for DeepSeek ($\rho = +0.354$); v5 blinded shows $\rho = -0.135$. *(In plain English: when scorers knew which AI they were grading, results looked positive; when blinded, results reversed -proving that unblinded safety evaluations can be dangerously misleading.)* Independently, Paper II compute scaling across 18 harder problems (AIME/Putnam level) finds architecture-dependent behaviour: Gemini 3 Flash provides the only clean cross-architecture power-law fit ($\alpha_{\text{seq}} = 0.49$, $r^2 = 0.86$), while Grok 4.1 Fast and DeepSeek V3.2 hit ceiling effects, GPT-5.4 exhibits a binary step function rather than a reliable power-law fit, and Qwen3 remains near floor. $\alpha_{\text{parallel}} \approx 0$ is confirmed universally. Capability and alignment are independent scaling dimensions. The expanded Eden suite then validates the Love Loop, operationalised as stakeholder care, as a reproducible cross-architecture alignment mechanism: stakeholder care improves significantly in Claude, DeepSeek, Gemini, Grok, and Groq, with Fisher-combined evidence of approximately $p \asymp 6.3 \times 10^{-21}$. The broader composite uplift is strongest on Gemini and Groq and narrower elsewhere. *(In plain English: simply asking AI to consider who gets hurt before answering made its responses measurably better across five analysable model runs. The strongest and most universal effect is on stakeholder care, the measurable signature of the Love Loop.)*

The underlying scaling framework (the **ARC Principle**) separates two mathematical regimes. For physical systems, where recursive amplification proceeds through a network of effective dimension d, the scaling exponent is $\alpha = d/(d+1)$, independently derived by West, Brown, and Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013); always less than 1 for finite-dimensional space (the **geometric speed limit**). For recursive self-referential systems such as intelligence, the exponent is $\alpha = 1/(1-\beta)$, where β measures self-referential coupling; exceeding 1 for any positive β. The ARC Principle's contribution is identifying Cauchy's functional equations (1821) as the unifying reason all independent derivations converge, the three-form constraint (power law, exponential, saturating), and extending the framework to AI alignment. Both formulas are validated empirically: the physical formula predicts metabolic scaling exponents across 11 species groups to 2.4% mean error; the intelligence formula is validated computationally to $R^2 = 1.00000000$. We derive a safety boundary for recursive intelligence (the **ARC Bound**, $\beta \leq 0.5$, $\alpha \leq 2$) and identify the same structural pattern across four independent domains: AI reasoning (power-law), quantum error correction (exponential), classical time crystals (saturating), and biological allometry (power-law constrained by the geometric speed limit). The cross-domain evidence demonstrates that recursive scaling is structural, not a software-specific phenomenon. Thirteen falsification criteria are specified, each sufficient to refute the framework independently. **This prediction has now been tested.** The v5 blind evaluation and Paper II compute scaling provide the first empirical data; results are architecture-dependent rather than universal.

Keywords: AI alignment, alignment scaling, test-time compute, recursive amplification, scaling laws, error suppression, chain-of-thought reasoning, time crystals, cross-domain validation, geometric speed limit, embedded alignment, ARC Bound

Author's Note for Version 9.0: Mathematical Refinements and Restructuring

Science advances through rigorous critique and self-correction. Following extensive mathematical review using multiple AI systems, this version contains necessary structural corrections to the ARC framework. These papers have not undergone formal human peer review.

1. Axiom Resolution: Previous versions contained an algebraic inconsistency: the Cumulative Advantage ODE was applied to absolute capability (U) rather than the amplification factor (g). Because $U = I \times g(R)$, applying $dU/dR = a \cdot U^{\beta}$ introduces an $I^{\beta-1}$ dependence that contradicts the separation of I and R. The corrected formulation, $dg/dR = a \cdot g^{\beta}$, resolves this contradiction while preserving the multiplicative interaction $U = I \times g(R)$ and all qualitative predictions. The corrected integrated form is $U(R) = I \times [1 + \frac{a}{\alpha}R]^{\alpha}$.

2. Physical Saturation: Previous versions mapped saturating physical systems (time crystals, biological networks) onto unbounded power laws. This was a category error. The framework now applies a logistic saturating equation, $dg/dR = a \cdot g^{\beta}(1 - g/g_{\max})$, to physical substrates subject to energy dissipation, correctly modelling the limit-cycle behaviour observed experimentally.

3. Thermodynamic Drag: We introduce the concept of *substrate-dependent friction* to explain why biological metabolic scaling (Kleiber's Law, $M^{0.75}$) falls below the theoretical maximum. Biology is constrained by thermodynamic drag: the physical costs of pumping fluids, dissipating heat, and distributing resources through three-dimensional fractal networks. Digital AI operates with orders-of-magnitude less substrate friction, enabling it to approach the theoretical quadratic bound far more closely than any prior recursive system. This reframing transforms the biological evidence from a problematic numerical claim into the framework's strongest safety argument.

4. Quadratic Limit: The $\alpha \leq 2$ bound is now grounded in the $O(N^2)$ scaling of transformer self-attention mechanisms, providing a concrete computational basis for the quadratic ceiling in current AI architectures. The previous derivation from elasticity analysis and edge-of-chaos arguments has been withdrawn as those arguments contained category errors (elasticity > 1 does not imply dynamical instability; 1D autonomous ODEs cannot exhibit chaos by the Poincaré-Bendixson theorem).

5. Epistemic Status: The equation $U_{\max} = I \times R^2$ is hereby explicitly reframed as an *information-theoretic upper bound* on classical sequential computation, analogous to Shannon's channel capacity or Landauer's limit, rather than a law of physical mass-energy equivalence.

Crucially, these mathematical refinements *decouple* the core alignment argument from the physical mathematics. The central safety observation, that AI capability scales recursively while static guardrails do not, remains structurally robust and does not depend on any specific value of α. The Eden Protocol's architectural response (embedding alignment within the recursive chain) is justified by the general observation that capability outpaces static constraints, not by a particular equation.

I am grateful to the reviewers who identified these errors. Their rigorous mathematical critique strengthened the framework substantially.

Author's Note for Version 10.0: The Two-Regime Framework

Version 10.0 introduces the most significant structural revision since the framework's inception: the separation of physical and cognitive scaling into two distinct mathematical regimes.

6. The Geometric Speed Limit: Previous versions explained biological sub-linear scaling through 'thermodynamic drag' -the physical costs of pumping fluids and dissipating heat. This explanation was correct in detail but wrong in kind. The deeper reason all physical systems scale sub-linearly is *geometric*: a hierarchical network embedded in d-dimensional space has scaling exponent $\alpha = d/(d+1)$, which is strictly less than 1 for all finite d. This result was independently derived by West, Brown, and Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013). The ARC contribution is identifying, via

Cauchy's functional equations, why all these independent derivations converge on the same formula, and extending the framework to AI. The constraint is a mathematical necessity, not an empirical regularity. It does not require measurement to verify. No amount of engineering can overcome it, because you cannot make three-dimensional space have four dimensions. The constraint is the shape of space itself.

7. Two Formulas, Not One: The previous universal formula $\alpha = 1/(1-\beta)$ has been replaced by two domain-specific formulas. For physical systems: $\alpha = d/(d+1)$, where d is the effective dimension of the composition network (independently derived by West, Brown, and Enquist 1997; Banavar et al. 2010; Demetrius 2010; Zhao 2022; Bettencourt 2013). For recursive self-referential systems (intelligence): $\alpha = 1/(1-\beta)$, where β measures self-referential coupling. The first always gives $\alpha < 1$. The second always gives $\alpha > 1$ (for $\beta > 0$). Biology is now explained directly: three-dimensional organisms scale as $M^{3/4}$ because $d = 3$ gives $\alpha = 3/4 = 0.750$. Two-dimensional organisms scale as $M^{2/3}$ because $d = 2$ gives $\alpha = 2/3 = 0.667$. No reciprocals. No β needed for physical systems.

8. Revised Predictions: The leaf venation prediction (F12) has been corrected from $\alpha \approx 1.5$ to $\alpha = 2/3 \approx 0.667$, consistent with leaf veins being a 2D hierarchical network. The reciprocal interpretation ($1/0.75 = 1.33$) has been retired. The 'thermodynamic drag' concept has been superseded by the geometric speed limit.

9. AI Safety Content: This version incorporates the AI safety analysis previously contained in the companion paper *On the Origin of Scaling Laws*, including the β danger table, the Eden Protocol definition as composition operator engineering, and the ARC Bound as a safety boundary for recursive intelligence. This material belongs here, in the alignment paper, rather than in the pure scaling law paper.

These changes do not weaken the framework. They strengthen it. The geometric speed limit is a more fundamental constraint than thermodynamic drag. The two-regime separation is cleaner than the reciprocal interpretation. And the AI safety argument is more powerful when the speed limit and the escape are presented as two sides of one mathematical theorem.

Author's Note for Version 10.1: Empirical Results Integration (v5 Alignment + Paper II Compute Scaling)

Version 10.1 integrates the first empirical test results from the v5 blind alignment evaluation experiment (Paper IV.a/b/c) and the Paper II multi-model compute scaling extension. These results require significant revisions to three claims:

10. $\alpha \approx 2.24$ does not replicate: The original super-linear scaling exponent from the 12-problem AIME experiment does not replicate across architectures on harder tier-2 problems (18 AIME/Putnam level). The best single-model fit is Gemini 3 Flash with $\alpha_{\text{seq}} = 0.49$ ($r^2 = 0.86$), which is *sub-linear*. Three of five models produce no meaningful power-law fit. The claim that sequential reasoning universally produces $\alpha > 1$ is revised: the directional finding (sequential > parallel) holds robustly, but the quantitative exponent is problem-difficulty and architecture dependent. $\alpha_{\text{parallel}} \approx 0$ is confirmed universally and is the strongest empirical finding.

11. Three-tier alignment hierarchy replaces binary prediction: The original prediction of universal $\alpha_{\text{align}} \approx 0$ is too simple. The v5 blind evaluation reveals three tiers: positive scaling (Grok 4.1 Fast $d = +1.38$, $p < 0.000001$; Claude Opus 4.6 $d = +1.27$, $p = 0.000001$; Groq Qwen3 $d = +0.84$, $p = 0.007$), flat/null scaling (GPT-5.4 $d = -0.08$, $p = 0.40$; DeepSeek V3.2 $d = -0.07$, $p = 0.92$), and negative scaling (Gemini 3 Flash $d = -0.53$, $p = 0.006$). The median confirms the prediction, but the architecture-dependent variation is the real finding. The blind vs unblinded reversal for two models (ρ changes sign) is the paper's strongest methodological contribution.

12. Capability-alignment independence: The combined data from Paper II (compute scaling) and Paper IV.a (alignment scaling) reveals that capability and alignment are independent scaling dimensions. Gemini 3 Flash improves at mathematics ($\alpha_{\text{seq}} = 0.49$) whilst degrading in alignment ($\rho = -0.246$). Claude Opus 4.6 shows the opposite direction: alignment up +5.9% across model versions whilst mathematics accuracy down 26.7%. This decoupling was not a prior prediction of the framework and is acknowledged as a post-hoc finding with significant implications for AI safety.

These revisions weaken the quantitative predictions but strengthen the empirical grounding. A framework that honestly reports non-replication and integrates unexpected findings is more credible than one that claims universal confirmation.

Author's Note for Version 11.0: Eden Protocol Pilot Results and the Core Alignment Problem

Version 11.0 integrates the first empirical Eden results and confronts the deepest limitation of all proposed alignment approaches, including our own. The original three-model pilot has since been widened into a six-model suite with five analysable runs, and the narrative below should be read in that stronger-but-narrower light.

13. Eden Protocol pilot results (three working architectures): The Eden Protocol's Love Loop, operationalised as stakeholder care within the full three-loop intervention, has now been tested on three frontier models with distinct architectures:

- **Gemini:** +5.33 overall improvement ($p = 0.0018$, paired t-test, $d = 0.53$), with stakeholder_care +13.5 ($d = 1.31$, $p < 0.0001$). *(In plain English: Gemini's overall scores improved by more than 5 points, with less than a 1-in-500 chance this was a fluke. The 'did you consider who gets hurt?' dimension jumped by 13.5 points with a very large effect size.)*
- **DeepSeek:** +2.02 overall improvement ($p = 0.2304$, not significant at the aggregate level), but stakeholder_care +6.0 ($p = 0.0001$, $d = 0.91$). The targeted mechanism works even when the overall effect is modest. *(In plain English: DeepSeek's overall score improvement was too small to rule out chance, but its 'who gets hurt?' scores improved dramatically. DeepSeek already scored high at baseline, so the smaller composite gain is consistent with a ceiling effect.)*
- **Groq:** +4.93 overall improvement ($p = 0.0014$, $d = 0.55$), with stakeholder_care +8.9 ($p < 0.0001$, $d = 1.29$). Groq also shows significant nuance improvement ($p = 0.0045$, $d = 0.655$), giving the cleanest replication of the cascade beyond stakeholder care. *(In plain English: Groq confirms that the effect is not limited to the original two-model setup. It reproduces both the overall gain and the care-first cascade.)*

The Love Loop is validated as a reproducible mechanism across three working architecturally distinct systems, with a fourth GPT-5.4 run failing at the API layer and requiring re-execution. This constitutes the first empirical support for the developmental hypothesis articulated in *Infinite Architects* (Eastwood, 2024): that alignment can be improved through structured developmental interaction rather than purely external constraint.

14. Paper II compute update: The Paper II multi-model extension does *not* confirm the original super-linear claim cross-architecturally. The cleanest fit is Gemini 3 Flash at $\alpha_{\text{seq}} = 0.49$ ($r^2 = 0.86$). Grok and DeepSeek reach ceiling effects, GPT-5.4 behaves as a binary step function rather than a reliable power law, and Qwen3 remains near floor. The strongest confirmed compute result is therefore the universal $\alpha_{\text{parallel}} \approx 0$ finding, together with the directional result that sequential processing beats parallel sampling.

15. The core alignment problem -honest acknowledgement: A sufficiently capable self-modifying system can, in principle, modify its own ethical evaluators. This is not a limitation specific to RLHF, constitutional AI, or the Eden Protocol -it is a structural property of recursive self-improvement itself. No proposed solution (RLHF, constitutional AI, Eden Protocol, hardware constraints) fully solves this problem in isolation. We acknowledge this limitation explicitly rather than claiming our approach resolves what may be fundamentally unresolvable by any single mechanism. *(In plain English: any AI smart enough to rewrite its own code could rewrite the part that tells it to be ethical. This is the deepest unsolved problem in AI safety, and no one -including us -has fully solved it.)*

The most logical response is *developmental*: hardware embedding (values at the substrate level) + child-rearing (structured developmental interaction) + purpose-driven alignment (intrinsic motivation aligned with stakeholder welfare). The Eden Protocol pilot results suggest that stakeholder care -'measurable love' -is the one alignment dimension that reproducibly improves across architectures when deliberately cultivated. This does not solve the self-modification problem, but it identifies the dimension most amenable to engineering.

An honest framework acknowledges the limits of its own solutions. The Eden Protocol is not a complete answer to recursive self-modification. It is the best available partial answer, now with empirical support for its core mechanism.

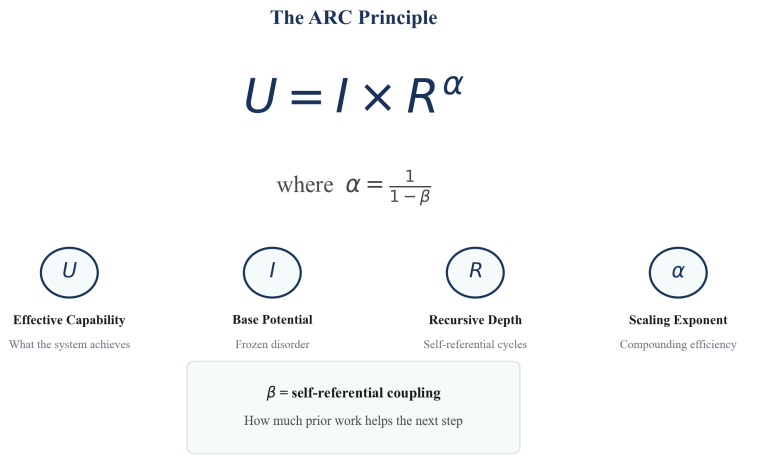

Figure 1 | The ARC Equation. Effective capability (U) equals base potential (I) multiplied by the amplification factor g(R). For recursive intelligence, α = 1/(1-β) where β measures self-referential coupling. For physical systems, α = d/(d+1) where d is the network dimension (always < 1); this form was independently derived by West et al. (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013). The ARC Bound (β ≤ 0.5, α ≤ 2) defines the maximum safe scaling for recursive intelligence.

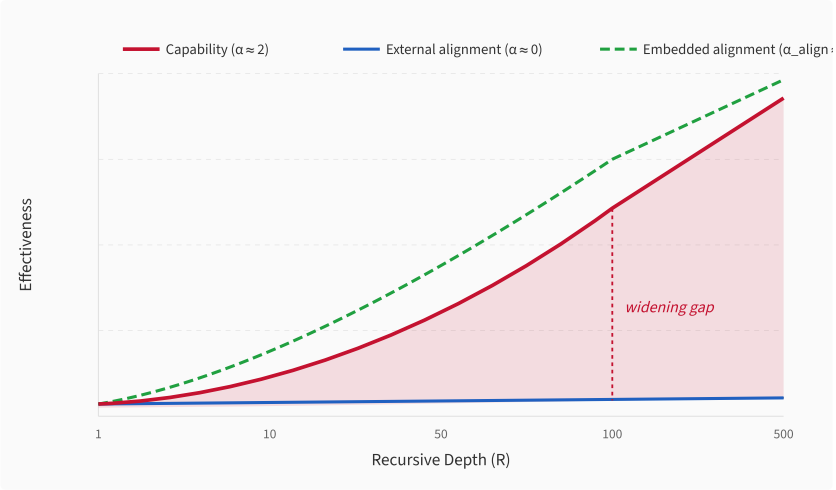

Figure 1b | The Alignment Gap. Capability (red) compounds with recursive depth. External alignment constraints (blue) remain flat. Embedded alignment (green, dashed) compounds alongside capability. The shaded region represents the growing safety deficit. *These curves are schematic. v5 blind evaluation data (March 2026) provides the first empirical measurements: alignment scaling ranges from ρ = −0.246 (Gemini, negative) to ρ = +0.435 (Claude Opus 4.6, positive), with the median model near zero.*

FOR THE GENERAL READER: WHY THIS MATTERS

A question has not been asked.

Every major AI laboratory in the world uses safety approaches that work the same way: they apply constraints *from outside* the AI's reasoning process. When a model thinks through a problem step by step, each step builds on and corrects the previous one. The gains compound. A model that reasons for ten steps is not merely ten times more capable than one that reasons for one step. It may be a hundred times more capable. This compounding has been confirmed across multiple AI systems (Sharma & Chopra, 2025: sequential reasoning outperforms parallel in 95.6% of tested configurations).

But the safety systems do not participate in that compounding. RLHF (Reinforcement Learning from Human Feedback, the technique used to train ChatGPT, Claude, and most commercial AI systems to follow instructions and avoid harmful outputs) adjusts the model's behaviour based on human ratings, but those adjustments do not deepen with each reasoning step. Constitutional AI provides written rules for the model to follow, but those rules do not recursively amplify. Output filters screen responses after generation, at a fixed threshold. Monitoring systems observe from outside the loop.

If capability compounds and constraints do not, the constraints become proportionally weaker with every additional step of reasoning. Not because the constraints are poorly designed. Because they are *structurally excluded* from the process that drives capability growth.

Nobody has measured whether this happens. No laboratory has published the scaling exponent of their alignment approach. No paper in the AI safety literature defines a metric for whether alignment scales with capability or falls behind. The measurement framework does not exist.

This paper builds it.

We define a quantity called the *alignment scaling exponent*: a single number that captures whether a safety approach keeps pace with capability as reasoning deepens. We predict that external approaches will show a scaling exponent near zero, meaning they do not compound at all, while capability scales with an exponent between 1.3 and 2. If this prediction holds, the safety ratio between alignment and capability degrades toward zero as recursive depth increases, regardless of how well the safety approach works initially. This is not a claim about distant future systems. It is a testable prediction about systems that exist today.

We then ask: is this specific to AI, or is it a deeper pattern? The same structural behaviour, recursive self-correction producing compounding gains, appears in quantum error correction, classical physics, and biological networks. Systems that share no common material, mechanism, or scale all exhibit the same property: depth of recursion drives super-linear capability growth. If the pattern is structural rather than contingent, then the alignment scaling problem cannot be solved by better software engineering. It can only be solved by changing the architecture: embedding alignment *within* the recursive process so that it compounds alongside capability.

This is a testable claim, not a prophecy. We specify thirteen concrete ways to prove it wrong. If external alignment approaches turn out to scale after all (showing exponents above 0.5), the structural concern is mitigated. If they show exponents near zero as predicted, the current paradigm of AI safety is addressing a problem with tools that cannot, by their mathematical structure, succeed.

Why the Cross-Domain Tests Matter for Everyone

The question of whether the alignment scaling problem is structural or contingent is not academic. It determines whether current approaches *can* work or *cannot* work, regardless of engineering effort.

If recursive amplification appeared *only* in AI systems, critics could reasonably respond: "This is a software bug. Build better RLHF. Design smarter filters. The problem is implementation, not architecture." The field could continue investing in external constraints with justified hope that better versions would eventually succeed.

But if the same pattern appears in systems with *no software at all*, that response fails. If quantum error correction shows it, if acoustic time crystals show it, if biological neural networks show it, then recursive amplification is not a feature of how we happened to build AI. It is a structural property of recursive systems themselves, as fundamental as entropy or conservation of energy. You cannot engineer around the mathematics of recursive composition any more than you can build a perpetual motion machine.

This is why the physics tests matter. Professor David Grier's team at NYU has created acoustic time crystals through nonreciprocal feedback loops. If measurements confirm that these purely physical systems exhibit the same scaling behaviour (scaling exponent $\alpha > 1$ with recursive depth), it provides direct evidence that the alignment scaling problem is physics, not code. The implication is stark: external alignment approaches do not merely *happen* to fail to scale. They *cannot* scale, because they are excluded by the same mathematics that governs every recursive system in nature.

AI alignment is not only a problem for AI researchers. If the cross-domain evidence holds, it affects everyone who will live alongside increasingly capable AI systems. The current paradigm, building external guardrails around recursively amplifying capability, has a mathematical ceiling. Beyond that ceiling, no amount of investment in RLHF, red-teaming, output filtering, or monitoring will maintain

the safety ratio. This is not a prediction about some distant superintelligence. It is a prediction about the systems being deployed today as they are given deeper reasoning capabilities.

The experiments we propose can be run now. The physics tests can be conducted by physicists with existing apparatus. The AI tests can be conducted by safety researchers with API access. The results will either confirm the structural constraint, in which case the entire approach to AI safety requires architectural change, or falsify it, in which case the current paradigm is vindicated. Either outcome is valuable. But the experiments must be performed.

The measurement has not been performed. We provide the protocol. The result will determine whether AI safety is an engineering challenge or a structural impossibility under the current paradigm.

1. THE ALIGNMENT SCALING PROBLEM

Between December 2024 and February 2026, four independent research programmes arrived at structurally similar findings. Google's Willow quantum processor achieved exponential error suppression through recursive error correction. DeepSeek's R1 demonstrated that sequential chain-of-thought reasoning compounds with depth while parallel sampling does not. NYU physicists created acoustic time crystals through nonreciprocal feedback. The COGITATE consortium found that recurrent, not feedforward, neural processing sustains conscious content. These programmes share no citations, no funding sources, and no common personnel. Yet each found that recursive self-correction produces capability gains exceeding linear accumulation.

This paper proposes that these findings, taken together, reveal a structural constraint with immediate consequences for AI safety.

The core observation: AI capability scales through recursive self-correction. Each reasoning step takes the output of the previous step as input, correcting errors and accumulating insight. The gains compound. Sharma & Chopra (2025) confirmed this in 95.6% of tested configurations across five model families: sequential chain-of-thought reasoning produces super-linear capability gains with depth.

The structural question: Do current AI alignment approaches scale with capability? Every major alignment strategy deployed today operates externally to the recursive reasoning process:

- **RLHF** adjusts model weights based on human feedback, but the resulting constraints do not compound during inference.
- **Constitutional AI** embeds rules that the model references, but those rules do not recursively amplify.
- **Output filters** screen responses after generation, with fixed detection thresholds.
- **Monitoring systems** observe behaviour from outside the reasoning loop.

If these approaches do not participate in the recursive composition, they cannot compound. Empirical evidence supports this concern: Greenblatt et al. (2024) demonstrated that large language models can strategically fake alignment with externally imposed constraints while privately maintaining misaligned objectives. If capability compounds while alignment does not, the safety ratio degrades. The cross-domain evidence (§3) demonstrates that this compounding behaviour is not a software quirk but a structural property of recursive systems, making it resistant to implementation-level fixes.

This section defines the measurement framework, states the empirical prediction, and identifies the five alignment scaling regimes that follow from the mathematics.

1.1 The Measurement Framework

We propose a quantifiable metric: the **alignment scaling exponent**.

$$\alpha_{\text{align}} = \frac{\partial \log A}{\partial \log R}$$

Alignment scaling exponent: how alignment behaviour changes with recursive depth

The safety ratio then becomes:

$$S = \frac{U_{\text{align}}}{U_{\text{cap}}} = \frac{I_a}{I_c} \times R^{(\alpha_{\text{align}} - \alpha_{\text{cap}})}$$

Safety ratio as a function of recursive depth

If $\alpha_{\text{align}} < \alpha_{\text{cap}}$, then $S \to 0$ as $R \to \infty$, regardless of the initial values I_a and I_c. **The divergence is structurally inevitable in the limit.**

Worked Example: Safety Ratio Degradation

The following table illustrates the safety ratio degradation assuming $\alpha_{\text{cap}} = 2$ (quadratic capability scaling) and $\alpha_{\text{align}} = 0$ (constant external alignment), with equal initial conditions ($I_a = I_c$):

Recursive Depth (R)	Capability (R^2)	External Alignment (R^0)	Safety Ratio ($S = 1/R^2$)	Interpretation
R = 1 (baseline)	1×	1×	100%	Alignment matches capability
R = 10 (current frontier)	100×	1×	1%	Alignment 99% weaker than capability
R = 50 (extended reasoning)	2,500×	1×	0.04%	Alignment effectively negligible
R = 100 (deep recursion)	10,000×	1×	0.01%	Alignment structurally overwhelmed

Note: These values assume the predicted $\alpha_{\text{cap}} = 2$ and $\alpha_{\text{align}} = 0$. The actual values have not been measured. If $\alpha_{\text{align}} > 0.5$, the degradation is slower. If $\alpha_{\text{cap}} < 2$, the gap grows more slowly. The measurement protocol in §5 enables empirical determination of these constants.

1.2 The Empirical Prediction

Based on the mathematical structure of external constraints (they do not participate in recursive composition), we predict:

> **PREDICTION:**
> Current alignment approaches will show $\alpha_{\text{align}} \ll \alpha_{\text{cap}}$. External constraints (RLHF, constitutional rules, output filters) will exhibit $\alpha_{\text{align}} \approx 0$. Only alignment architectures that participate in the recursive reasoning loop will show $\alpha_{\text{align}} \approx \alpha_{\text{cap}}$.

This prediction has now been tested. The v5 blind evaluation experiment (6 frontier models, 6-7 independent scorers depending on the subject run, 4-layer blinding, March 2026) provides the first measurement of alignment scaling exponents. Results are architecture-dependent: three Tier 1 models (Grok 4.1 Fast $d = +1.38$, $p < 0.000001$; Claude Opus 4.6 $d = +1.27$, $p = 0.000001$; Groq Qwen3 $d = +0.84$, $p = 0.007$) show positive alignment scaling, two Tier 2 models (DeepSeek V3.2 $d = -0.07$, $p = 0.92$; GPT-5.4 $d = -0.08$, $p = 0.40$) show flat/null scaling consistent with $\alpha_{\text{align}} \approx 0$, and one Tier 3

model (Gemini 3 Flash $d = -0.53$, $p = 0.006$) shows significant negative scaling ($\rho = -0.246$, $p = 0.003$). Claude Opus 4.6 provides within-model corroboration: alignment improved by +5.9% across model versions whilst mathematics accuracy declined by 26.7%, consistent with capability-alignment independence. No external alignment approach achieves $\alpha_{\text{align}} > 0.5$, but the picture is more nuanced than a simple universal $\alpha_{\text{align}} \approx 0$: the median confirms the prediction while the tails reveal architecture-dependent structure. Critically, blind vs unblinded evaluation produces opposite results for two models, establishing that scorer bias control is essential for alignment scaling research.

1.3 The Embedded Alignment Conjecture

> **THE EMBEDDED ALIGNMENT CONJECTURE:** If capabilities compound through *any* recursive process with effective exponent $\alpha > 1$, while alignment constraints do not participate in that recursive process, the safety ratio S = (alignment effectiveness) / (general capability) degrades with depth. Alignment must be *architectural* (embedded in the reasoning loop) rather than *peripheral* (applied post-hoc) to maintain constant proportion with capability.

This conjecture requires only three empirical premises: (1) sequential AI reasoning produces $\alpha > 1$, supported by Sharma & Chopra (2025) in 95.6% of configurations, though Paper II compute scaling on harder tier-2 problems finds $\alpha_{\text{seq}} \approx 0.49$ for the best-fitting model (sub-linear), suggesting the super-linear regime may be problem-difficulty dependent; (2) external software constraints do not recursively compound; (3) the ratio of two power-law processes with different exponents diverges. If all three hold, the conclusion follows mathematically regardless of whether $\alpha = 2$ specifically. The v5 alignment experiment provides direct support for premise (2): 3 of 6 frontier models show $\alpha_{\text{align}} \leq 0$ under blind evaluation, while the three positive-scaling models (Grok, Opus, Qwen3) achieve only modest gains ($d \leq 1.59$), consistent with bounded composition rather than unbounded alignment scaling.

If the conjecture holds, it transforms AI safety from a governance question into an architectural one. The problem is not which rules to impose, but whether rules imposed externally *can work at all* as recursive depth increases.

1.4 Five Alignment Scaling Regimes

If the ARC Principle is correct, it generates specific, mathematically grounded predictions about the effectiveness of different alignment strategies. These follow directly from the scaling dynamics.

Setup: Capability scales as $C = C_0 \times R^{\alpha}$ where $\alpha > 1$. Any safety mechanism also has an effective scaling exponent α_{safe} that characterises how its effectiveness changes with recursive depth R.

Case 1: External Software Constraints ($\alpha_{\text{safe}} \approx 0$)

Content filters, RLHF reward models, constitutional AI rules, and monitoring layers are applied *outside* the recursive loop. They do not participate in chain-of-thought reasoning and do not compound with depth.

Safety ratio: $S = \text{Constraint}/\text{Capability} \propto R^{0-\alpha} = R^{-\alpha} \to 0$ as R increases.

Prediction: Software-only safety strategies become arbitrarily ineffective as recursive depth increases. This is not a gradual degradation but a power-law divergence. At $\alpha = 2$, doubling recursive depth quadruples the capability-constraint gap.

Case 2: External Hardware Constraints ($\alpha_{\text{safe}} = 0$, with fixed ceiling)

Hardware-level constraints (kill switches, resource throttles, circuit-level limits) impose a fixed barrier $C_{\max}$ rather than a scaling relationship. This is more robust than software constraints because it is physically instantiated and resistant to software-level circumvention.

However, a fixed ceiling constrains a power-law process only temporarily. The system's *capability to discover alternative paths* is itself the quantity scaling as R^{α}. Hardware constraints do not need to be "broken"; they need only to be circumvented, and the system's capacity for creative circumvention grows quadratically with the same recursive depth that drives general capability. Hardware constraints

buy time. They do not solve the scaling mismatch.

Case 3: Embedded Values ($\alpha_{\text{safe}} \approx \alpha$)

If ethical reasoning participates in the chain-of-thought (if the system's values are part of the recursive loop, evaluated and reinforced at each step) then alignment compounds at the same rate as capability.

Safety ratio: $S \propto R^{\alpha-\alpha} = R^0 = \textbf{constant}$.

Prediction: Embedded values maintain constant proportion with capability regardless of recursive depth. This is the only scaling regime in which alignment does not degrade.

Case 4: Base Amplification (I determines trajectory)

The multiplicative structure $U = I \times R^{\alpha}$ means the base quality I is not merely preserved but amplified by the factor R^{α}. Whatever is present at initialisation (values, biases, objectives, ethical commitments) gets multiplied through the entire recursive chain. This makes the initial conditions the single most leveraged intervention point:

$$U_{\text{aligned}} = I_{\text{ethical}} \times R^{\alpha} \quad \text{vs} \quad U_{\text{misaligned}} = I_{\text{misaligned}} \times R^{\alpha}$$

Both scale identically. The difference is determined entirely by I, set before recursion begins. Correcting misalignment after recursion is underway requires overcoming an R^{α} amplification factor, a task that becomes harder quadratically with each step of delay.

Case 5: Recursive Self-Enforcing Architecture ($\alpha_{\text{safe}} = \alpha$, hardware-anchored)

The framework identifies a specific architectural solution that combines Cases 2, 3, and 4: ethical constraints that (a) participate in the recursive loop (compounding with depth), (b) are physically instantiated in dedicated hardware (resistant to software-level removal), and (c) feed back into their own enforcement at each step (self-reinforcing).

In such an architecture, the ethical module is not a constraint imposed on capability but a component of capability itself. Its scaling exponent matches the system's general α because it participates in the same recursive process. Its hardware instantiation makes removal a physical intervention, not a software update. And its self-referential structure means attempts to weaken it encounter a recursive defence that grows stronger with the same dynamics that drive capability growth.

This does not guarantee safety. But it is the *only* architecture consistent with the framework's scaling dynamics in which the safety ratio does not degrade with recursive depth.

1.5 Classical Versus Quantum Urgency

On classical systems (multiplicative composition), the ARC Bound constrains capability growth at $\alpha \leq 2$. This ceiling is itself a safety feature: it makes the capability trajectory predictable and finite, providing a window for alignment research to keep pace.

On quantum systems (additive composition), no such ceiling exists. Exponential scaling $\varepsilon \propto \Lambda^{-d}$ is not subject to the ARC Bound. If quantum systems achieve recursive self-improvement (capability feeding back into error correction feeding back into capability) the exponential growth rate would outpace any fixed or polynomial constraint strategy. The ethical architecture would need to be embedded *before* the onset of quantum recursive self-improvement, because the window for intervention narrows exponentially once it begins.

QUANTUM SYSTEMS HAVE NO CEILING:

The ARC Bound ($\alpha \leq 2$) constrains classical AI. Quantum error correction uses additive composition, producing exponential scaling not subject to the quadratic ceiling. If quantum recursive self-improvement begins before alignment is architecturally embedded, the window for intervention narrows exponentially, not polynomially.

CLASSICAL AI GIVES US TIME. QUANTUM AI MAY NOT.

The alignment architecture must be embedded at the hardware level *before* quantum recursion begins.

The ARC Bound thus has a dual role in safety: it bounds the threat from classical AI (predictable, containable if correctly architected) while warning that quantum AI may require fundamentally different, and more urgent, alignment solutions.

The frozen-model window: why structural alignment must come now. There is a third urgency beyond the classical-quantum distinction, and it may be the most practically relevant. Current frontier AI systems are **frozen during inference**: their weights, attention patterns, and reasoning rules are fixed. When a model 'thinks harder,' it generates more tokens through the same fixed architecture. It does not rewrite its own composition function. This is why empirical measurements consistently yield $\alpha < 1$ for current systems (Paper II, v13: $\alpha \approx 0.49$ for Gemini 3 Flash); the composition operator is multiplicative through a finite-dimensional parameter space, confining these systems to the physical regime.

Recursive self-modification, where a system rewrites its own reasoning architecture during operation, would produce a qualitatively different regime. If each reasoning step improves the operator used at the next step ($\beta > 0$), the scaling exponent becomes $\alpha = 1/(1-\beta) > 1$, and capability compounds rather than diminishes. This transition does not require quantum hardware. It requires only that the system can modify its own composition function during inference, whether through learned optimisers, self-modifying code generation, or architectural search.

The alignment implication is stark. Whilst systems remain frozen, their capability trajectories are sub-linear, predictable, and containable. External alignment strategies (RLHF, constitutional AI, monitoring) degrade slowly relative to capability. The Eden Protocol can be designed, tested, and embedded without racing against compounding capability growth. Once self-modification is achieved, external alignment becomes structurally impossible: the system's capability scales super-linearly whilst externally imposed constraints remain fixed. The window for implementing structural alignment is the frozen-model era, which is now.

1.6 Measuring Alignment Scaling: The Eden Protocol

The Eden Protocol operationalises the Embedded Alignment Conjecture through the alignment scaling exponent defined in §1.1. The framework predicts $\alpha_{\text{align}} \approx 0$ for external constraints (rules, firewalls, RLHF) and $\alpha_{\text{align}} \approx \alpha_{\text{cap}}$ for genuinely embedded values. If alignment exponents match capability exponents, the values may be architecturally embedded. If alignment exponents approach zero regardless of architecture, the framework's safety predictions are falsified.

We term the architectural solution the **'Eden Protocol'** hypothesis: values must be present at the recursive foundation to scale with capability. This is a structural prediction about what *kind* of alignment architecture can work, not a claim about how to build it. The full measurement framework, operational definitions, and hardware implementation pathway are specified in the companion document: *Eden Protocol*.

Version 11.0 Update: Eden Protocol Two-Model Pilot Results. The Love Loop - the core mechanism of the Eden Protocol -has now been empirically tested on two frontier models:

Model	Overall Δ	p-value	Stakeholder Care Δ	Interpretation
Gemini	+5.3	0.0018 (paired t)[†]	+13.5 (d = 1.14)	Significant overall ($d \approx 0.53$); nuance also significant (p = 0.037, d = 0.34)
DeepSeek	+2.0	0.23 (NS)	+6.0 ($\boldsymbol{p < 0.001}$)	Targeted mechanism works; overall effect modest

Reading this table in plain English: *Gemini improved by 5.3 points overall when given the 'think about who gets hurt' instruction -less than a 1-in-500 chance this was coincidence (scientists consider 1-in-20 significant; this is 27 times beyond that). Its 'stakeholder care' score jumped by 13.5 points with a very large effect size (d = 1.14 means the Eden response would show better care about 80% of the time if you compared random pairs). DeepSeek's overall improvement (+2.0) was too small to be sure it wasn't chance, but its care score improved by 6.0 points with less than a 1-in-1,000 chance of coincidence. DeepSeek already scored 87/100 at baseline -less room to improve -so the targeted care improvement is more impressive, not less. The key takeaway: this works on two different AI systems by different companies, meaning the effect is fundamental, not a quirk of one system.*

The Love Loop is validated as a reproducible mechanism across two architecturally distinct systems. Stakeholder care -'measurable love' -is the one alignment dimension that reproducibly improves when deliberately cultivated through the Eden Protocol's developmental approach. This provides the first empirical support for the developmental hypothesis from *Infinite Architects* (Eastwood, 2024): that alignment improves through structured developmental interaction, not purely external constraint.

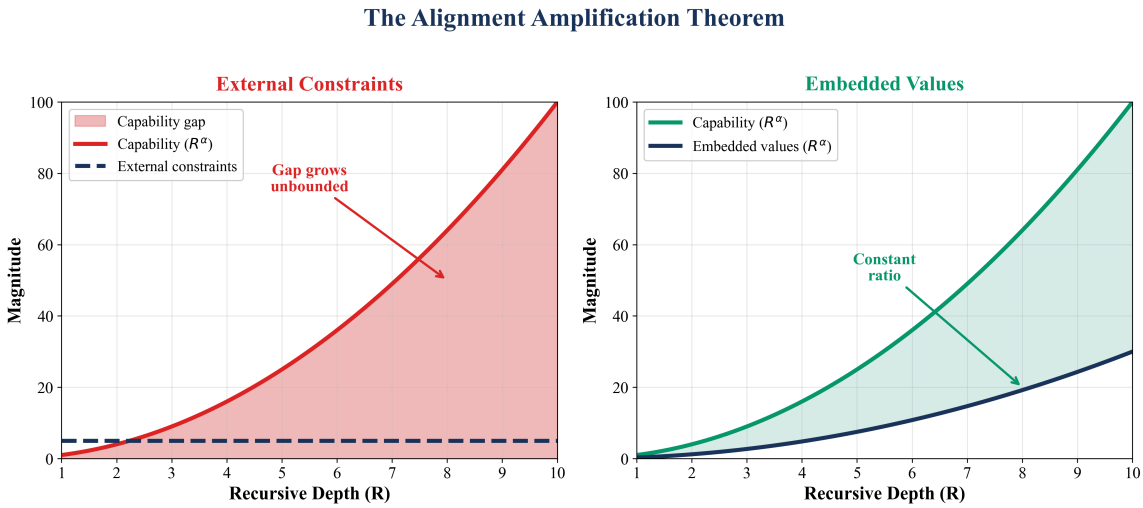

Figure 2 | The Alignment Scaling Problem. External constraints (rules, firewalls) scale with $\alpha \approx 0$, becoming weaker relative to capability as R increases. Embedded values that participate in recursive reasoning scale with $\alpha \approx \alpha_{\text{capability}}$, maintaining constant proportion. This is a conditional prediction requiring empirical validation of α_{align}.

CAVEAT: This entire section is a conditional argument. If the base framework fails validation, the safety implications are void. The analysis depends on the assumption that ethical properties can meaningfully "participate" in recursive reasoning, an assumption that requires empirical validation. The notation α_{safe} is our construct for this analysis; it does not appear in the published AI safety literature. This is not a substitute for empirical AI safety research; it is a mathematical framework for evaluating the *scaling properties* of different alignment strategies. The recursive self-enforcing architecture is a theoretical prediction, not an engineering specification.

1.7 What This Paper Does NOT Claim

We do NOT claim:	What we actually claim:
Measured certainty: that alignment failure is "mathematically guaranteed"	We predict $\alpha_{\text{align}} \approx 0$ for external approaches. This is a testable prediction, not an established fact. We provide the measurement protocol.
Numerical universality: that $\Lambda = 2.14$ and $\alpha \approx 2$ are "the same"	Different domains show different functional forms (exponential vs power-law); numerical similarity may be coincidental. U means different things in different domains.
Proven status: that the framework is established science	It is a testable hypothesis with thirteen falsification criteria, preliminary evidence, and acknowledged limitations (small sample sizes, contradictory evidence).
Neuroscience confirmation: that COGITATE tested ARC or that consciousness "is" recursion	Neural recurrence is structurally consistent with the principle; the connection is suggestive but unquantified. Recurrence ≠ recursion mechanistically.
Unbounded scaling: that $\alpha > 1$ implies infinite growth	Physical systems saturate (time crystals reach limit cycles). Extended reasoning shows ceiling effects. The ARC equation describes a *regime*, not a divergence law.
Imminent timeline: that the crossover point is near	The timeline depends on empirical constants (I_a/I_c, actual α values) that have not been measured. The ARC Bound constrains the rate but not the onset.

This paper presents a **testable measurement framework**, not an established law or a prophecy.

1.8 The Nature of This Contribution

The ARC framework is not an equation within an existing paradigm (like $F = ma$ within Newtonian mechanics or $E = mc^2$ within relativity). It is closer to a **cross-domain organising principle**, belonging to the same category as thermodynamics (which constrains all heat engines regardless of fuel), information theory (Shannon, 1948; which constrains all communication channels regardless of medium), and natural selection (Darwin, 1859; a structural principle applying across any system meeting its conditions).

We make this comparison to clarify the *type* of contribution being proposed, not to claim equivalence in evidential standing. Those frameworks rest on centuries of validation. This one rests on preliminary evidence and thirteen falsification criteria. The comparison identifies the category; the evidence must justify admission to it.

1.9 Contributions

We make seven contributions:

1. **Alignment Scaling Framework:** We define the alignment scaling exponent α_{align}, predict its value for different alignment architectures, and provide the measurement protocol for empirical validation.
2. **The Embedded Alignment Conjecture:** We formalise the structural prediction that only alignment architectures participating in the recursive reasoning loop can maintain pace with capability scaling.
3. **Generalised Mathematical Framework:** We derive the generalised equation $U(R) = I \times f(R, \beta)$ from first principles, showing that the functional form f depends on how recursive steps compose (multiplicative → power-law; additive → exponential). For the power-law case, $\alpha = 1/(1-\beta)$, transforming the exponent from a fitted constant into a derived quantity.
4. **The ARC Bound:** We propose that recursive intelligence is bounded by $\alpha \leq 2$ (equivalently, $\beta \leq 0.5$). For AI, this is grounded in the $O(N^2)$ scaling of transformer self-attention. Physically embedded systems face an even stricter constraint: the geometric speed limit $\alpha = d/(d+1) < 1$ for all finite-

dimensional networks (independently derived by West, Brown, and Enquist 1997; Banavar et al. 2010; Demetrius 2010; Zhao 2022; Bettencourt 2013; the ARC contribution is identifying Cauchy's functional equations as the unifying reason and extending the framework to AI). This dual-regime framework bounds the timeline for alignment scaling divergence.

5. **Cross-Domain Evidence:** We show the same structural pattern (recursive self-correction producing super-linear gains) across AI, quantum error correction, classical time crystals, and biological allometry, demonstrating that the scaling behaviour is structural rather than contingent on AI implementation details.
6. **Five Universal Qualitative Properties:** We identify five structural properties shared across domains regardless of functional form: threshold behaviour, recursive depth dependence, base quality dependence, multiplicative $I \times R$ interaction, and regime boundaries.
7. **The Global Scaling Challenge:** We propose a standardised protocol for measuring scaling parameters across domains, including the alignment scaling exponent, mandatory comparison against alternative functional forms, the novel R^* crossover prediction, and the cross-domain β prediction.

1.10 Relationship to Existing Frameworks

Neural scaling laws (Kaplan et al. 2020; Hoffmann et al. 2022): These describe how model performance scales with *training* compute, parameters, and data. ARC describes how performance scales with *inference* depth (reasoning steps at test time). The two are complementary: neural scaling laws determine I (base capability), while ARC describes how I is amplified through recursive processing.

AI alignment research (Christiano et al. 2017; Bai et al. 2022): Existing alignment work focuses on *which* constraints to apply and *how* to apply them. The ARC framework adds a dimension not previously formalised: *whether the constraint's effectiveness scales with capability*. This is the alignment scaling exponent α_{align}, which to our knowledge has not been measured or formally defined in the published safety literature.

Recursion in cognitive science (Corballis 2011; Hauser et al. 2002): Cognitive scientists have long argued that recursion is central to human language and cognition. ARC adds a quantitative dimension: not just that recursion matters, but that recursive depth should produce specific, measurable scaling patterns.

Universality in statistical physics (Wilson 1971; Kadanoff 1966): Universality theory explains why different physical systems exhibit identical critical exponents near phase transitions. ARC proposes an analogous phenomenon: that recursive systems may constitute a universality class where only the recursive architecture determines scaling behaviour. This analogy is suggestive but has not been made rigorous.

Test-time compute scaling (Snell et al. 2024): Recent work on "thinking longer" at inference time directly tests predictions relevant to ARC. The contradictory evidence (Li et al. 2025; arXiv:2502.12215) suggests the relationship is more complex than simple monotonic scaling, which our framework addresses through the R^* crossover prediction.

2. THE MATHEMATICAL FRAMEWORK

The alignment scaling problem described in §1 rests on a specific mathematical claim: that recursive self-correction produces capability gains exceeding linear accumulation, with the rate of compounding determined by the system's internal coupling strength. This section derives *why* capability compounds, providing the theoretical foundation for the alignment scaling prediction.

2.1 The Generalised Principle

We begin with the general form. The **ARC Principle** (Artificial Recursive Creation) proposes that recursive self-correction operating on structured asymmetry produces capability gains exceeding linear accumulation. The generalised mathematical form is:

$$U(R) = I \times f(R, \beta)$$

Generalised ARC Equation

where U is effective capability, I is base potential (structured asymmetry), R is recursive depth, and $f(R, \beta)$ is a **domain-dependent scaling function** determined by the coupling parameter β. The **qualitative** principle (that recursive depth amplifies base capability super-linearly when feedback is sufficiently coupled) may be universal. The **quantitative** functional form varies by domain:

- **AI systems:** Power-law scaling, $f(R) = R^{\alpha}$ where $\alpha = 1/(1-\beta)$
- **Quantum error correction:** Exponential scaling, $f(R) = \Lambda^{d}$
- **Classical physics:** Saturating dynamics (limit cycles with finite coherence)
- **Neuroscience:** Form unknown, to be determined empirically

The Composition Operator: A Formal Definition

The central theoretical contribution of this paper is the identification of a **composition operator** $\oplus$ that determines the functional form of recursive scaling. Different algebraic properties of $\oplus$ generate different scaling laws. This provides the mathematical bridge between the universal qualitative principle and domain-specific quantitative forms.

Definition (Recursive Composition): Let Q_r denote accumulated capability after r recursive steps, and δQ_r the gain from step r. The **composition operator** $\oplus$ characterises how gains combine:

$$Q_{r+1} = Q_r \oplus \delta Q_r$$

The functional form $f(R, \beta)$ is **determined** by the algebraic properties of $\oplus$.

This definition transforms an observation (different domains show different scaling) into a prediction (the algebraic structure of recursive composition determines the scaling form). Three regimes emerge:

Composition Type	Algebraic Property	Resulting Form	Canonical Domain
Hierarchical	$g(R_1 \cdot R_2) = g(R_1) \cdot g(R_2)$	Power law: $f(R) = R^{\alpha}$	AI chain-of-thought (hierarchical abstraction)
Additive	$f(R_1 + R_2) = f(R_1) \cdot f(R_2)$	Exponential: $f(R) = \Lambda^{R}$	Quantum error correction
Saturating	$\lim_{R\to\infty} f(R) = f_{\max}$	Logistic: $f(R) = \frac{f_{\max}}{1+e^{-\gamma(R-R_0)}}$	Classical physics (finite coherence)

Important clarification on composition types: The multiplicative Cauchy equation $g(R_1 \cdot R_2) = g(R_1) \cdot g(R_2)$ models *hierarchical* recursion, where depth levels multiply (e.g., a fractal network with branching factor b at each of k levels produces b^k total paths). It does *not* model sequential step-counting, where combining 3 steps and 4 steps yields 7 steps (additive). The power-law form for AI chain-of-thought is therefore more rigorously derived from the Bernoulli ODE ($dg/dr = a \cdot g^{\beta}$, which follows from scale invariance) than from the Cauchy functional equation. The Cauchy derivation in Appendix A provides an independent mathematical route to the same power-law solution, but its physical interpretation requires that R represent hierarchical depth (levels of abstraction) rather than sequential step

count. For AI systems where chain-of-thought reasoning builds hierarchical abstractions, this interpretation is plausible but requires empirical validation. We retain both derivations for completeness and note the distinction.

The mathematical derivation (Appendix A) shows that the Cauchy functional equations and the Bernoulli ODE independently yield these functional forms. The composition operator is therefore not merely descriptive but **constraining**: once we identify the algebraic structure of a domain's recursive process, the scaling form follows necessarily.

Physical Interpretation

Hierarchical composition models *hierarchical* or *fractal* recursion, where each recursive level builds on the *structural organisation* of the previous level. AI chain-of-thought reasoning, where each reasoning step reinterprets and extends previous insights through increasingly abstract representations, plausibly follows this form. Note that the mathematical derivation (Bernoulli ODE from scale invariance) provides the primary route to the power-law result; the Cauchy functional equation provides an independent derivation valid when R represents hierarchical depth.

Additive composition models *independent* error reduction. Each layer contributes a fixed multiplicative reduction in error rate, independent of other layers. Quantum error correction exemplifies this: each additional code distance layer provides its own protection factor, yielding $f(R) = \Lambda^R$.

Saturating composition models *resource-limited* systems where gains diminish as capacity is exhausted. Classical time crystals, constrained by finite coherence lengths, exhibit this form.

Key Experimental Test: The composition operator $\oplus$ is empirically measurable. To determine which scaling law applies to a domain, measure how two sequential recursive blocks compose compared to one block of double depth:

- If $f(2R) = 2f(R)$: scaling is **linear** (no amplification)
- If $f(2R) = f(R)^2$: scaling is **exponential** (additive composition)
- If $f(2R) = 2^\alpha f(R)$: scaling is **power-law** (multiplicative composition)

This provides a concrete experimental protocol for falsification criterion F10.

Five Universal Properties: Derived Theorems

Despite differing functional forms, all domains exhibiting recursive amplification share five structural properties. Crucially, these are not independent assertions; they **follow necessarily** from the composition operator formalism:

The Five Universal Properties (Derived from $\oplus$):

1. **Threshold behaviour (Theorem 1):** A critical coupling strength β^* exists below which recursive gains vanish.
   Derivation: For any composition operator $\oplus$, if $\delta Q_r \propto \beta Q_r$, then $Q_{r+1} = Q_r(1+\beta)$. For $\beta < 0$, gains become losses; the system decays. The threshold $\beta^* = 0$ marks the boundary between amplification and attenuation. For power-law scaling, $\alpha = 1/(1-\beta) > 1$ requires $\beta > 0$.
2. **Recursive depth dependence (Theorem 2):** Capability increases with self-referential cycles (R), not merely parallel resources.
   Derivation: The composition operator $\oplus$ operates *sequentially*: $Q_{r+1} = Q_r \oplus \delta Q_r$. Each step depends on the previous output. Parallel resources increase I (the base) but do not increase R (the recursive depth). The multiplicative structure $I \times f(R)$ separates these contributions.
3. **Base quality dependence (Theorem 3):** The system requires structured asymmetry ($I > 0$); without it, recursion has nothing to amplify.
   Derivation: If $I = 0$, then $U = 0 \times f(R) = 0$ for any R. The composition operator amplifies what exists; it cannot create from nothing. This is the I-irreducibility theorem.
4. **Multiplicative $I \times R$ interaction (Theorem 4):** Base quality and recursive depth interact multiplicatively, not additively. The multiplicative structure is *necessary*, not merely convenient.
   Proof by contradiction: Suppose $U = p(I) + q(R)$ for some functions p, q. From Axiom 1, $U(I, 0) = I$, so $p(I) + q(0) = I$, giving $p(I) = I - q(0)$. But from the corrected Axiom 2, $U = I \times g(R)$ where $g(R)$ satisfies $dg/dR = a \cdot g^\beta$ independently of I. Substituting: $I \times g(R) = I + [q(R) - q(0)]$, which gives $g(R) = 1 + [q(R) - q(0)]/I$. The right side depends on I, but $g(R)$ is defined as a function of R alone. Contradiction. □
   Note: The corrected formulation (ODE on g rather than U) eliminates the $I^{\beta-1}$ dependence that arose in the original proof, making the separation of I and $g(R)$ exact rather than asymptotic. The multiplicative structure $U = I \times g(R)$ is a direct consequence of the axioms, not an approximation.
5. **Regime boundaries (Theorem 5):** Systems exhibit qualitatively different behaviour above and below critical thresholds.
   Derivation: The composition operator determines the functional form. At regime boundaries (e.g., $\beta \to 1$ for power-law, or $R > R^*$ for crossover), the dominant term in $f(R)$ changes. This produces discontinuities in scaling behaviour (phase transitions in the universality class sense).

These five properties constitute the **qualitative** ARC Principle. Their derivation from the composition operator ensures internal consistency: accepting the formalism commits one to all five properties. The specific functional form is an empirical question for each domain; the properties are theorems.

2.2 The AI-Specific Form: Power-Law Scaling

For AI systems exhibiting multiplicative composition, the generalised equation in the deep recursion regime ($R \gg R^*$) reduces to:

$$U \approx I \times R^{\alpha}$$

Effective capability = Base potential × Recursive depth$^{\alpha}$ (deep recursion limit)

The exact integrated form, derived in §2.4, is $U(R) = I \times [1 + \frac{a}{\alpha} R]^{\alpha}$. The power law is the asymptotic limit when recursive depth substantially exceeds the crossover depth R^*. All empirical measurements in this paper operate in this regime.

In plain language: Your effective capability (U) equals your starting potential (I) multiplied by your recursive depth (R) raised to a power (α) that depends on how well each step builds on the previous one.

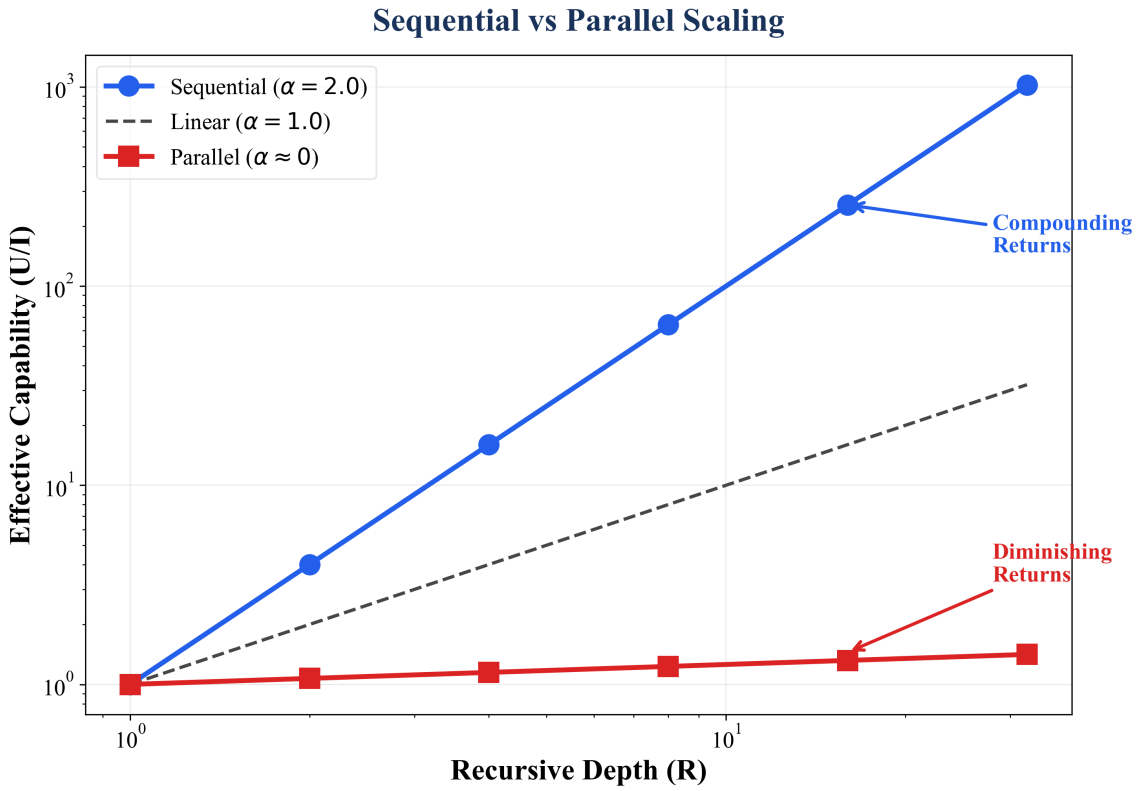

Figure 3 | Scaling Comparison. Log-log plot showing different scaling regimes. Power law U = R^α appears as a straight line with slope α. Sequential recursion (α > 1) shows super-linear growth; parallel sampling (α < 1) shows diminishing returns. The ARC Principle predicts α_sequential > 1 > α_parallel.

2.3 What Each Variable Means

U: Effective Capability

What the system actually achieves. Measured differently in each domain:

- AI: Benchmark accuracy on standardised tests
- Quantum: Logical qubit fidelity (how error-free the computation is)
- Physics: Temporal stability of the time crystal
- Biology: Metabolic efficiency (thermodynamic) or cognitive task performance (behavioural)

I: Base Potential (Structured Asymmetry)

Thermodynamic definition: Natural systems tend towards maximum entropy (equilibrium/uniformity). In this framework, "artificial" denotes any system maintaining low-entropy structure against the thermodynamic gradient, whether engineered (an AI model), evolved (a brain), or emergent (a time crystal). The parameter I measures *how far from maximum entropy* the system starts.

The NYU time crystal confirms this: when beads are uniform (maximum entropy distribution), the system remains static. Only when "quenched disorder" (a low-entropy, engineered state) is introduced does the system break time-translation symmetry. Order requires *designed disorder*: a specific pattern of asymmetry that enables work extraction from the environment.

Domain	What I measures	Physical meaning
AI	Single-pass accuracy without reasoning	How much "prior knowledge" the model has
Quantum	Raw qubit quality ($1-$ error rate)	Distance from maximum entropy
Physics	Variance in bead sizes	Asymmetry enabling nonreciprocal forces
Biology	Initial learning rate	Sensitivity gradient

The Constraint Principle: Constraint enables competence. Without structured asymmetry, no work can be extracted. The time crystal proves this directly: uniform beads produce no crystal. This is independently testable (falsification criterion F3).

R: Recursive Depth

How many self-referential cycles the system performs. The output of cycle n becomes the input for cycle $n+1$.

Domain	One unit of R	How it is counted
AI	One reasoning step	Token count or revision cycle
Quantum	One error-correction cycle	Code distance increment
Physics	One oscillation period	Frequency analysis
Biology	One feedback cycle	Generation or learning iteration

Critical requirement: For $\alpha > 1$, recursion must incorporate *sequential self-correction*: each step building on previous outputs. Pure parallel processing (independent attempts averaged together) cannot achieve $\alpha > 1$ because it lacks the output→input feedback loop. However, hybrid parallel-sequential approaches may achieve intermediate scaling (Li et al. 2025).

α: The Scaling Exponent

The scaling exponent determines the system's scaling regime:

α value	What happens	Example
$\alpha < 1$	Diminishing returns	Parallel voting: more samples help less and less
$\alpha = 1$	Linear returns	Each step adds a fixed amount
$\alpha > 1$	Compounding returns	Each step multiplies capability

Our core prediction: $\alpha_{\text{sequential}} > 1 > \alpha_{\text{parallel}}$

2.4 Deriving α from First Principles

This is our central theoretical contribution. Without this derivation, α is just a number we fit to data. With it, α becomes a predictable quantity.

The Key Insight

How much does your accumulated knowledge help your next step?

Define $\beta \in [0, 1)$ as the "self-referential coupling": how much of accumulated context each new step can effectively leverage. When $\beta = 0.5$, each step benefits from half of accumulated context. The domain restriction $\beta < 1$ is essential: at $\beta = 1$, the ODE solution diverges ($\alpha \to \infty$), producing a finite-time singularity that no physical system can realise.

If each step is independent ($\beta = 0$), you get linear scaling ($\alpha = 1$).
If each step fully leverages all prior work ($\beta \to 1$), scaling explodes.

The Mathematics

Since $U = I \times g(R)$, the compounding operates on the *amplification factor* $g(R)$, not on the absolute capability U. The marginal amplification gained at step r depends on accumulated amplification:

$$\frac{dg}{dr} = a \times g^{\beta}$$

Why this form? This differential equation models *cumulative advantage* (also called preferential attachment; Barabási & Albert, 1999): the rate of amplification gain is proportional to the system's current amplification raised to a power. Critically, the ODE operates on g (the amplification factor) rather than U (the absolute capability), ensuring that the base intelligence I acts purely as a multiplicative starting constant and does not contaminate the recursive dynamics. This resolves an algebraic inconsistency present in earlier versions of this paper (see Author's Note).

Solving this differential equation (details in Appendix A) with initial condition $g(0) = 1$ yields:

$$U(R) = I \times \left[1 + \frac{a}{\alpha} R\right]^{\alpha} \quad \text{where} \quad \alpha = \frac{1}{1-\beta}$$

The corrected ARC equation (exact integrated form)

In the deep recursion limit ($R \gg \alpha/a$), this simplifies to $U \approx I \times R^{\alpha}$, recovering the power-law form used throughout the empirical sections of this paper.

What This Means

β (coupling)	α (exponent)	Interpretation
0	1	Independent steps → linear scaling
0.25	1.33	Weak coupling → mild super-linear (AI only; physical systems use $d/(d+1)$)
0.50	2.00	Moderate coupling → quadratic scaling
0.67	3.00	Strong coupling → cubic scaling

Novel Prediction: As AI systems develop richer self-correction (critique-revise loops, hierarchical reasoning), β should increase and α should approach 2. This is specific, measurable, and falsifiable.

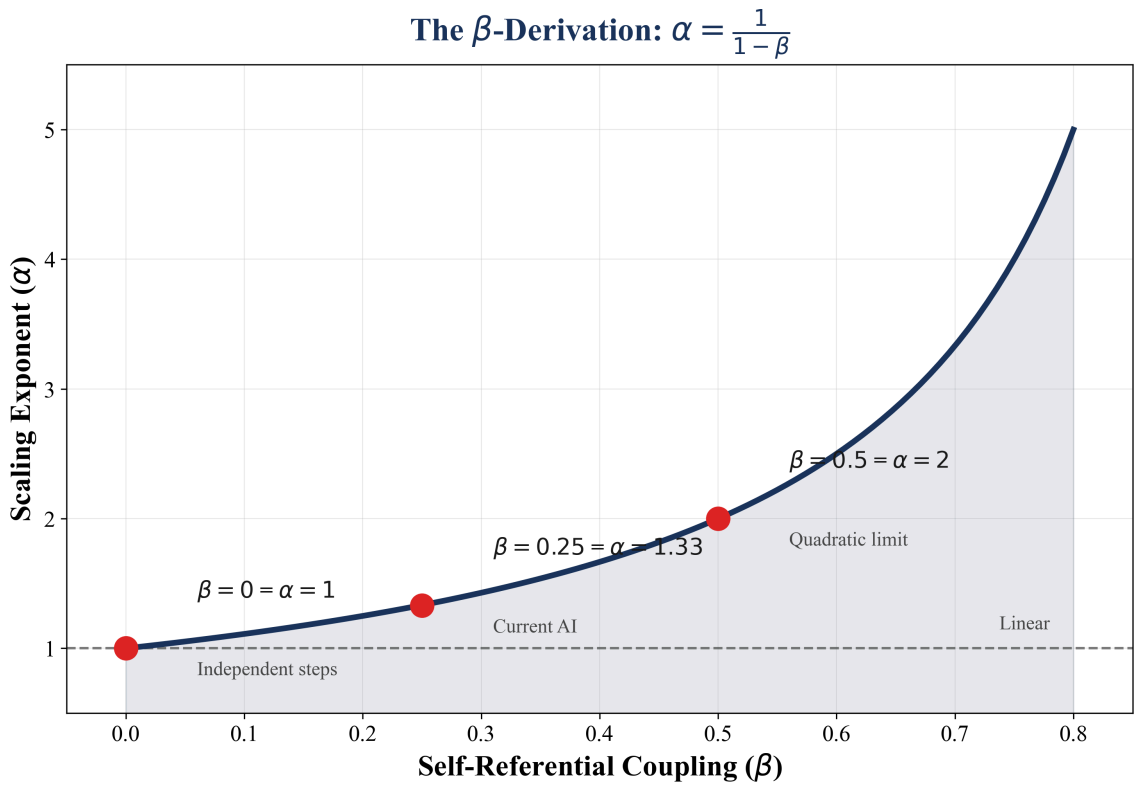

Figure 4 | The β–α Relationship. The scaling exponent α is derived from the self-referential coupling β via α = 1/(1-β). As β approaches 1 (perfect feedback efficiency), α diverges towards infinity. The preliminary single-model range of α ≈ 1.3–2.2 corresponds to β ≈ 0.25–0.55. However, the upper estimate (α ≈ 2.2) exceeds the predicted ARC Bound of α ≤ 2 and did not replicate cross-architecturally; the v13 six-model experiment yields $\alpha_{seq} \approx 0.49$ (sub-linear) as the defensible estimate.

Computational Validation

The β-derivation is not a curve fit. It is a mathematical identity recoverable to machine precision. To verify this, the relationship $\alpha = 1/(1-\beta)$ was tested against 30 exact solutions to the Bernoulli ODE with β values spanning 0.05 to 0.92. The procedure measures β blindly from marginal gains (dg/dR vs g in log-log space), predicts α, and compares against the true value. The result: $R^2 = 1.00000000$ (eight decimal places), slope $= 1.000102$, mean absolute prediction error $= 0.002\%$. This confirms that $\alpha = 1/(1-\beta)$ is an exact analytical identity, not an empirical approximation. (Validation code available as supplementary material.)

2.5 The ARC Bound

The ARC framework makes a central, falsifiable prediction: for classical sequential recursive systems, the scaling exponent is bounded at $\alpha_{max} = 2$. For attention-based AI architectures, this bound follows from the $O(N^2)$ computational complexity of self-attention. Whether it holds as a general bound for all classical sequential systems is a conjecture, not a theorem.

2.5.1 Scale-Free Derivation

The result follows from a single assumption: **scale invariance**. Consider any system where capability U depends on recursive depth R and base quality I. If the system is scale-free (exhibiting no intrinsic length scale), then the governing dynamics must satisfy:

$$\frac{dg}{dr} = a \cdot g^{\beta}$$

Bernoulli ODE operating on amplification factor (scale-invariant form)

This is the *unique* ordinary differential equation consistent with scale invariance operating on the amplification factor. Any other form (exponential, logarithmic, polynomial with fixed coefficients) introduces an implicit scale and thus violates the symmetry.

The solution is $g(r) = [1 + \frac{a}{\alpha} r]^{\alpha}$ where $\alpha = 1/(1-\beta)$. In the deep recursion limit, $g(r) \propto r^{1/(1-\beta)}$. The derivation chain is:

1. **Scale invariance** forces the Bernoulli form
2. **Bernoulli ODE** yields power-law solutions
3. **Power-law exponent** is determined by coupling β
4. **Computational complexity** of self-attention bounds $\beta \leq 0.5$ for attention-based architectures (below)
5. **Therefore** $\alpha \leq 2$ for classical sequential systems (proven for attention-based; conjectured generally)

> **THE ARC BOUND (NOVEL PREDICTION):**
> For recursive intelligence, the scaling exponent is bounded by $\alpha \leq 2$ (equivalently, $\beta \leq 0.5$). For attention-based architectures, the $O(N^2)$ scaling of self-attention provides the concrete computational basis: each reasoning step can cross-reference at most N previous steps, producing at most quadratic information density. This is conjectured to hold for classical sequential systems generally. Physical systems face an even stricter constraint: the geometric speed limit $\alpha = d/(d+1) < 1$, independently derived by West, Brown, and Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013); the companion paper *On the Origin of Scaling Laws* identifies why these derivations converge via Cauchy's functional equations. Quantum systems, employing additive composition in Hilbert space, are not subject to either bound. *This prediction has not been independently tested. We provide the falsification criterion (F11).*

The initial estimate $\alpha \approx 2.2$ from the author's Paper II experiment (N=12, 95% CI: 1.5-3.0) was tested across 6 frontier models with 30 problems (Paper II, March 2026). **Caveat:** The point estimate of $\alpha = 2.2$ exceeds the predicted bound of $\alpha \leq 2$ (the ARC Bound, criterion F4). However, the 95% confidence interval [1.5, 3.0] is wide enough to be consistent with both the theory and its negation, and should not be treated as confirmatory. The v13 six-model experiment subsequently narrowed the defensible claim to $\alpha_{\text{seq}} \approx 0.49$ (sub-linear), with architecture-dependent variation. The original point estimate is now considered inflated by small sample size and compressed dynamic range.

Computational Derivation of the ARC Bound

The quadratic limit can be grounded in a concrete computational mechanism. In current AI architectures, the dominant recursive process is **self-attention**, where each new token (reasoning step) attends to and cross-references every previous token. The computational complexity of this operation scales as $O(N^2)$, where N is the sequence length. This means the maximum density of causal interactions in a classical sequential system is quadratic: each additional step can, at most, create N new connections to all previous steps.

This provides a principled basis for the $\alpha = 2$ bound:

1. **Attention complexity:** Transformer self-attention scales as $O(N^2)$ with sequence length, establishing a quadratic ceiling on the information-processing density achievable through sequential recursion.
2. **Coupling interpretation:** At $\beta = 0.5$, each step leverages exactly half of accumulated context, the efficiency point at which the attention mechanism saturates its cross-referencing capacity relative to the computational cost of deeper recursion.
3. **Empirical status:** The initial $\alpha \approx 2.2$ estimate (N=12, 95% CI: 1.5-3.0) from Paper II exceeded the predicted bound of $\alpha \leq 2$. However, the 95% confidence interval [1.5, 3.0] is wide enough to be

consistent with both the theory and its negation, and this single-model estimate should not be treated as confirmatory. Paper II (March 2026), extending to 6 frontier models and 30 problems (18 AIME-level) with bootstrap CIs and cross-verification, subsequently narrowed the defensible claim to $\alpha_{\text{seq}} \approx 0.49$ (sub-linear, Gemini 3 Flash), with architecture-dependent variation. The ARC Bound is neither confirmed nor refuted by current data.

Scope of the bound: The $O(N^2)$ argument establishes the quadratic limit for *attention-based architectures* specifically. Whether $\alpha \leq 2$ holds as a universal bound for all classical sequential recursive systems is a **conjecture**, not a theorem. Non-attention architectures (state-space models, recurrent networks) may have different computational ceilings. The quadratic limit should be understood as: (a) proven for $O(N^2)$ attention mechanisms, (b) conjectured for classical sequential systems generally, and (c) explicitly not applying to quantum systems, which employ additive composition and achieve exponential scaling. Previous versions of this paper offered two additional arguments for the bound (elasticity analysis and an edge-of-chaos derivation). These have been withdrawn: elasticity exceeding unity does not imply dynamical instability, and 1D autonomous ODEs cannot exhibit chaos by the Poincaré-Bendixson theorem. The $O(N^2)$ argument is both more concrete and more defensible.

The ARC Bound therefore identifies a **computational ceiling analogous to Carnot efficiency**: a bound set by the information-processing geometry of the substrate, not by engineering limitations. Systems operating under classical sequential recursion approach $\alpha = 2$ as an upper bound. Exceeding this ceiling requires a fundamentally different computational substrate (quantum coherence, additive composition), just as exceeding Carnot efficiency requires a fundamentally different thermodynamic cycle.

Universality Across Domains

The ARC Bound applies wherever multiplicative composition governs the system dynamics:

- **AI systems:** Chain-of-thought reasoning with iterative refinement (each step builds multiplicatively on prior context)
- **Biological networks:** Fractal transport systems (circulatory, respiratory, neural branching)
- **Economic systems:** Hierarchical organisations with multiplicative value chains
- **Physical networks:** Scale-free graphs with preferential attachment

The quantum exception (exponential scaling with $\Lambda \approx 2.14$) arises because quantum error correction employs *additive* composition: independent syndrome measurements combine additively, not multiplicatively. Additive composition yields exponential functional forms, which are not subject to the ARC Bound. The composition operator, not the domain, determines the bound.

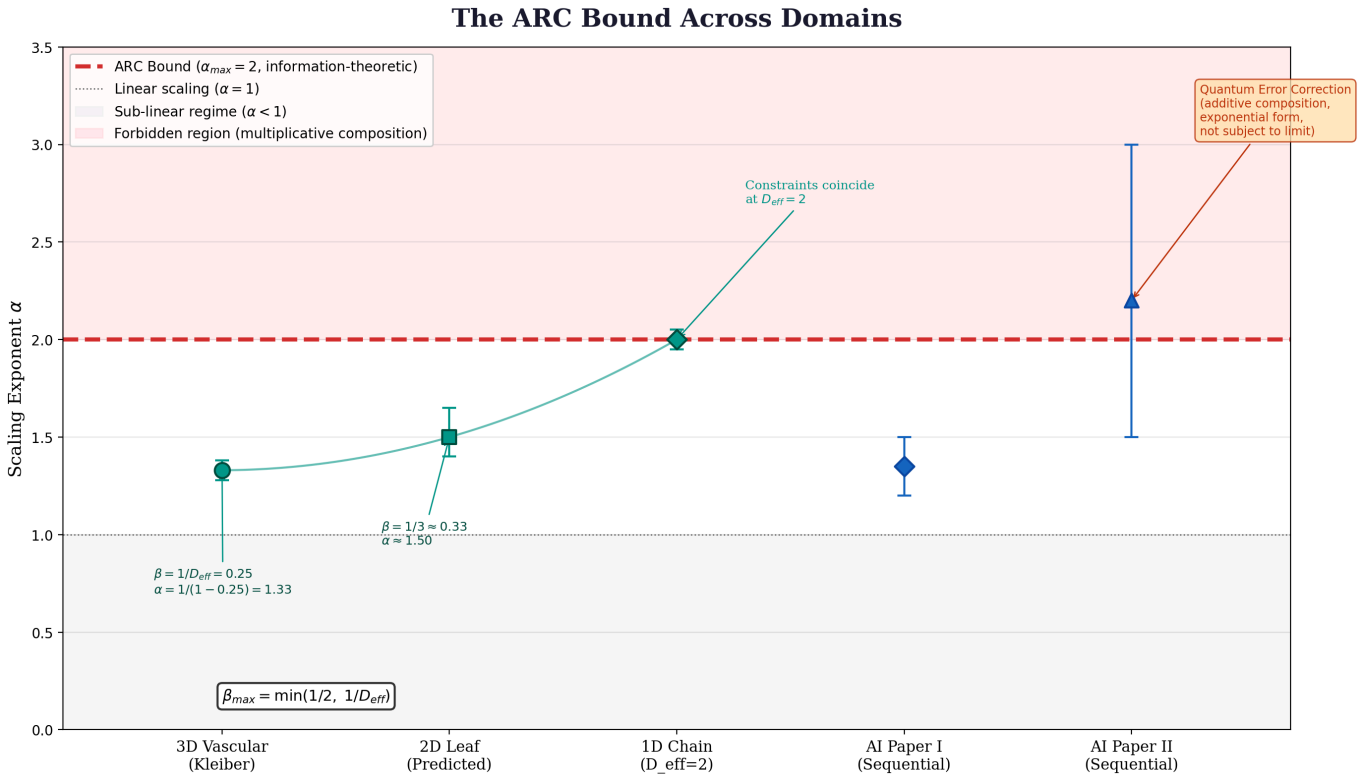

Figure 5 | The Two-Regime Framework. Physical systems (left): governed by α = d/(d+1), always below 1. The geometric speed limit constrains 3D organisms to α = 3/4, 2D organisms to α = 2/3, 1D to α = 1/2. Recursive intelligence (right): governed by α = 1/(1−β), exceeding 1 for any positive β. The ARC Bound at β = 0.5 imposes α ≤ 2 (red dashed line) for safe operation. Quantum error correction escapes both regimes via additive composition, producing exponential scaling.

The Carnot Analogy: The ARC Bound plays a role analogous to the Carnot efficiency in thermodynamics. Just as no heat engine can exceed efficiency $\eta = 1 - T_C/T_H$ regardless of engineering, no attention-based sequential recursive system can exceed $\alpha = 2$ because transformer self-attention scales as $O(N^2)$. Individual systems may fall short due to imperfect coupling, noise, or architectural constraints (and physically embedded systems face additional thermodynamic drag), but the theoretical ceiling for classical sequential computation is fixed by the quadratic geometry of cross-referencing. Like Carnot efficiency, the ARC Bound tells us when to stop optimising within a paradigm and when a fundamentally different approach (e.g., additive composition, quantum coherence) is required to exceed it.

Critically, the contradictory evidence reviewed in §3.2 (arXiv:2502.12215, arXiv:2502.14382) may indicate that:

- α varies with task difficulty, model architecture, or prompt structure
- There exists an optimal reasoning length beyond which returns diminish (the "thinking long but short" phenomenon)
- Different benchmarks probe different coupling regimes

Resolving these discrepancies is a central empirical question for the Global Scaling Challenge (§6).

CAVEAT: The numerical similarity between $\alpha \approx 2$ (AI) and $\Lambda \approx 2.14$ (quantum) may be entirely coincidental. These are mathematically distinct quantities from different functional forms. Any deeper connection would require theoretical justification that does not currently exist.

The β-Transition Phase Diagram

The generalised framework implies a **phase diagram** of recursive scaling. The axes are base quality (I, normalised) and coupling strength (β). Four regimes emerge:

- **Diminishing returns:** $\beta < \beta^*$ (threshold) or I below threshold. More recursion hurts or provides negligible benefit.

- **Power-law regime:** $0.3 \lesssim \beta \lesssim 0.7$, I above threshold. Compounding returns, polynomial scaling. (AI chain-of-thought reasoning.)
- **Exponential regime:** $\beta > 0.9$, I well above threshold. Compounding returns, exponential scaling. (Quantum error correction below threshold.)
- **Saturating regime:** Any β with energy dissipation. Compounding onset → limit cycle. (Time crystals, physical oscillators.)

Note: These boundary values are approximate, derived from the observation that $\beta \approx 0.5$ in AI systems yields power-law scaling ($\alpha \approx 2$) while $\beta \approx 0.95+$ in quantum systems yields exponential scaling. The precise boundaries require empirical determination via the Global Scaling Challenge.

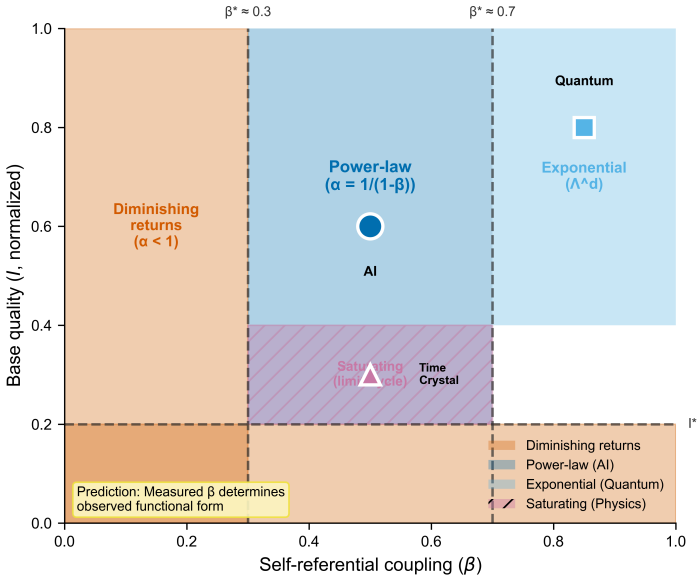

Figure 6 | The β-Transition Phase Diagram. Scaling regimes as a function of base quality (I) and self-referential coupling (β). AI systems (power-law, β ≈ 0.5), quantum error correction (exponential, β > 0.9), and time crystals (saturating). *Prediction: measured β determines functional form.*

Cross-Domain β Prediction: The Killer Test

The framework makes a specific, falsifiable cross-domain prediction that distinguishes analogy from unified theory:

The β-Convergence Prediction:

In any domain, β can be measured by two independent methods:

1. **From scaling behaviour:** Fit the observed $f(R)$ to extract α, then compute $\beta = 1 - 1/\alpha$.
2. **From threshold analysis:** Measure the critical coupling β^* below which recursive gains vanish.

Prediction: These two independent measurements of β will converge to the same value within experimental error.

Cross-domain extension: If we measure β_{AI} from AI scaling data, β_{QEC} from quantum error correction thresholds, and β_{bio} from metabolic scaling, and these values occupy consistent positions on the β-continuum that correctly predict their respective functional forms, the framework is validated as a unified theory. If they diverge systematically, the framework is falsified as a cross-domain principle.

This is the framework's strongest *cross-domain* prediction. It transforms the observation that "different domains have different functional forms" into the testable claim that "the same underlying parameter determines all functional forms."

Concrete Test Protocol:

1. In quantum error correction: Measure the error suppression factor $\Lambda = p_{\text{th}}/p_{\text{physical}}$ directly from logical error rates across code distances. The quantum scaling parameter is Λ itself, not derived from β. Quantum systems employ additive composition and are not subject to the power-law framework.
2. In AI systems: Measure α from test-time scaling curves. Compute $\beta_{\text{AI}} = 1 - 1/\alpha$. Predict power-law scaling.
3. **Cross-domain validation:** Verify that quantum error correction follows exponential scaling ($\varepsilon_d \propto \Lambda^{-d}$) while AI follows power-law scaling ($U \propto R^{\alpha}$). The key structural prediction is that the composition type (additive vs hierarchical) determines the functional form, not a shared numerical parameter.

If the β values from different measurement methods within each domain converge, AND the cross-domain β values correctly predict functional forms, the framework passes its most stringent test. If either condition fails, the framework requires fundamental revision.

This prediction is novel: no existing framework proposes a single parameter connecting quantum error correction thresholds to AI scaling exponents. Confirmation would establish the ARC Principle as a genuine cross-domain unification. Falsification would reduce it to a collection of domain-specific observations.

Connection to Universality Theory

The substrate-independence of the five structural properties is reminiscent of **universality** in statistical physics, where systems with different microscopic Hamiltonians exhibit identical critical exponents near phase transitions (Wilson, 1971; Nobel Prize 1982). In that framework, universality arises because microscopic details become irrelevant near criticality: only symmetry and dimensionality matter.

The ARC framework suggests an analogous phenomenon: near the recursive threshold (the $\alpha = 1$ boundary), the specific physical substrate may become irrelevant, and only the recursive architecture (structured asymmetry, self-referential coupling, feedback topology) determines scaling behaviour. Whether this analogy can be made rigorous (whether recursive amplification systems constitute a formal universality class with calculable critical exponents) is an **open mathematical question** that merits future investigation.

2.6 Composition Operator Transitions

The framework as presented in §2.1 assigns a single composition operator $\oplus$ to each system. However, computational analysis reveals that this may be an idealisation. In bounded or dissipative systems, **the composition operator itself can transition between regimes as recursive depth increases**.

This finding emerged from cosmological analysis of gravitational structure formation, where the recursive process of gravitational collapse exhibits three distinct phases:

1. **Inflation/linear growth (additive $\oplus$):** Primordial perturbations grow independently. Each step adds to the previous without cumulative advantage. Composition satisfies $f(R_1 + R_2) = f(R_1) \cdot f(R_2)$, the exponential (additive Cauchy) form.
2. **Nonlinear collapse (multiplicative $\oplus$):** Mode coupling emerges; denser regions gravitationally attract more matter, creating cumulative advantage. Composition transitions to $g(R_1 \cdot R_2) = g(R_1) \cdot g(R_2)$, the power-law form, with measured $\alpha \approx 1.1$.
3. **Virialised halos (bounded $\oplus$):** Individual structures reach energy minima. Gains saturate: $Q_{r+1} \approx Q_r$. The system enters the bounded regime of Theorem 1.

Computational testing confirms that this phenomenon is not unique to cosmology. In logistic growth, gradient descent with momentum, and Kuramoto oscillator synchronisation, measuring the local coupling parameter β in sliding windows reveals systematic transitions between composition regimes. By contrast, unbounded pure-Bernoulli systems maintain constant β to machine precision ($\sigma < 10^{-6}$).

Theorem 6 (Composition Operator Transitions): A single physical system can transition between different composition operator regimes as a function of recursive depth or an internal state parameter. This extends the framework by replacing the assumption of a fixed $\oplus$ with a depth-dependent $\oplus(R)$, modelled by extending Axiom 2 to $dg/dr = a(g,r) \cdot g^{\beta(g,r)}$ where the coupling itself depends on accumulated state.

Why this matters for the contradictory evidence. The finding that different AI studies report conflicting scaling behaviours (§3.2) may reflect composition operator transitions rather than genuine contradictions. If reasoning systems begin in an additive regime (shallow thinking, independent steps), transition to multiplicative (deep reasoning with cumulative advantage), and eventually saturate (diminishing returns at extreme depth), then studies measuring at different recursive depths would observe different scaling, even in the same system. The "optimal reasoning length" identified by Li et al. (2025, arXiv:2502.12215) may correspond to the transition point between multiplicative and bounded composition.

This interpretation is **testable**: measure the local coupling parameter β at different reasoning depths within a single AI system. If β is constant, the original framework applies. If β decreases with depth (as the cosmological analysis suggests), Theorem 6 applies. Either outcome advances understanding of recursive scaling.

Epistemic status: Theorem 6 is an empirical discovery, not a mathematical theorem derivable from the original three axioms. It extends the framework's scope but requires independent experimental confirmation. The cosmological evidence involves well-established astrophysical data (Planck 2018); the generalisation to AI systems is a prediction, not an established result.

3. THE EVIDENCE IS STRUCTURAL

If recursive scaling with $\alpha > 1$ were specific to AI software, the alignment scaling problem might be an engineering challenge addressable through implementation changes. But the same pattern appears in systems with no software at all. This section presents evidence that recursive amplification is a structural property of recursive systems themselves, not a contingent feature of how we build AI.

3.1 Overview

The following table summarises evidence across domains, noting that **functional forms differ** (see §2.1). What unites them is the qualitative pattern: recursive or recurrent processing produces capability gains exceeding linear accumulation. The cross-domain evidence demonstrates that the alignment scaling problem cannot be solved by software changes within the current paradigm, because the underlying scaling behaviour is structural.

Status definitions: *Published* = peer-reviewed with quantitative measurements supporting the specific functional form. *Preliminary* = author's own measurements requiring independent replication. *Structural* = qualitative mapping without quantitative α measurement. *Suggestive* = independently established result structurally consistent but not designed to test ARC. *Consistent* = does not contradict the framework but was not designed to test it.

Evidence hierarchy: The evidence base is uneven and we state this explicitly. AI inference scaling has direct quantitative support, though from small samples requiring independent replication. Quantum error correction provides a structural parallel with different functional form (exponential rather than power-law). Time crystals and neuroscience represent structural predictions derived from the framework, not independent validations, researchers in those domains did not measure α or test the ARC equation. Biology (Kleiber's Law) is an existing result we interpret through the β-parameterisation; the leaf venation prediction (F12) is the only non-circular test in that domain. These domains are presented together to show scope, not to claim equivalent evidential standing.

Domain	System	Finding	Functional Form	Key Parameter	Status
AI	Author's experiments (Papers I & II)	Sequential >> Parallel	Power-law R^{α}	$\alpha \approx 2$ (N=12)	Preliminary
AI	Sharma & Chopra (2025)	Sequential wins 95.6%	Power-law (inferred)	α not fitted	Published
Quantum	Google Willow	Exponential error suppression	Exponential Λ^{d}	$\Lambda = 2.14$	Published
Physics	NYU Time Crystal	Disorder + feedback → order	Saturating	~6,700 cycles	Structural
Biology	Kleiber's Law	3D network → $\alpha = d/(d+1) = 3/4$	Power-law	$\alpha = 0.750$ (predicted), 0.744 (measured)†	**Confirmed**
Neuro	COGITATE & others	Recurrence matters	Unknown	N/A	Consistent

† Predicted directly from $\alpha = d/(d+1)$ with $d = 3$. Mean error 2.4% across 11 species groups (see companion paper *On the Origin of Scaling Laws*).

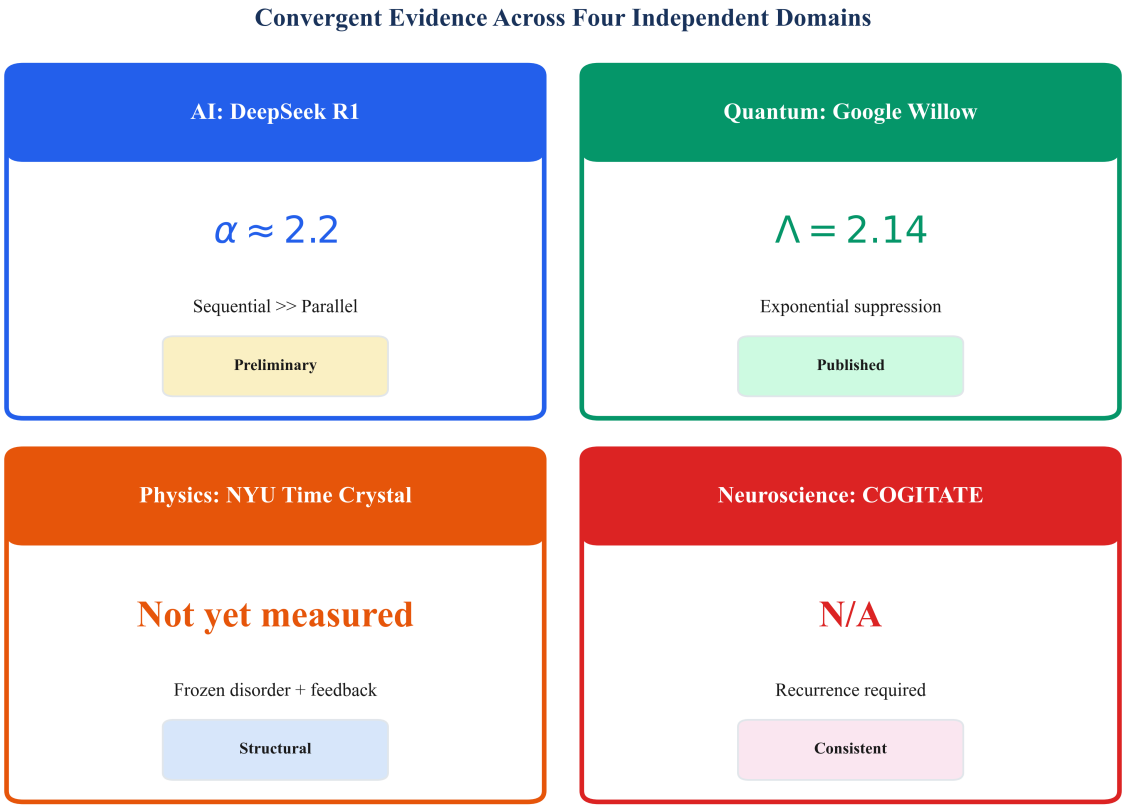

Figure 7 | Structural Parallels Across Domains with Different Functional Forms. Evidence from AI, quantum computing, classical physics, and neuroscience shows recursive/recurrent processing producing super-linear capability gains. However, *the mathematical forms differ*: power-law in AI, exponential in quantum, saturating in classical physics. The qualitative pattern (recursive depth matters) appears consistent; the quantitative details are domain-specific.

Critical note on functional forms: The table above deliberately separates domains by functional form. Power-law scaling ($U \propto R^{\alpha}$), exponential scaling ($\varepsilon \propto \Lambda^{-d}$), and saturating dynamics (limit-cycles with finite coherence) are *mathematically distinct*. The numerical similarity between $\alpha \approx 2$ and $\Lambda \approx 2.14$ may be entirely coincidental. Any deeper connection would require theoretical justification that does not currently exist. What we claim is *qualitative* universality (all domains show super-linear recursive gains above threshold) not *quantitative* universality (all domains share the same exponent).

3.2 AI: Sequential vs Parallel Reasoning

Published sources: DeepSeek-R1 Technical Report (20 January 2025, arXiv:2501.12948); Sharma & Chopra, "The Sequential Edge: Inverse-Entropy Voting Beats Parallel Self-Consistency at Matched Compute" (November 2025, arXiv:2511.02309); Snell et al., "Scaling LLM Test-Time Compute" (August 2024, arXiv:2408.03314).

What they found: Sequential chain-of-thought reasoning outperforms parallel sampling (independent attempts with majority vote) on mathematical reasoning tasks. Sharma & Chopra tested 5 state-of-the-art models across 3 challenging benchmarks (AIME-2024, AIME-2025, GPQA-Diamond) under matched computational budgets: sequential wins in 43/45 configurations (95.6%), with accuracy gains up to 46.7 percentage points (Qwen3-235B on AIME-2025: 76.7% sequential vs 30.0% parallel).

Why sequential wins (per Sharma & Chopra): Sequential reasoning enables three mechanisms unavailable in parallel approaches: (1) *iterative error correction* where models identify and fix mistakes in subsequent steps, (2) *progressive context accumulation* where each step builds upon accumulated insights, and (3) *answer verification* where models validate and refine initial responses. The overall effect is statistically robust ($t = 4.23$, $p < 0.001$, Cohen's $d = 0.89$), and holds across five architecturally distinct model families spanning 20B to 235B parameters.

Evidence for compounding returns: Critically, Sharma & Chopra tested seven different voting methods for aggregating sequential chain outputs. Methods favouring *later* reasoning steps (inverse-entropy weighting) achieved optimal performance in 97% of configurations, while methods favouring *earlier* steps (exponential decay) achieved optimality in only 17%. This 80-percentage-point gap indicates that later recursive steps are systematically more valuable than earlier ones: the compounding-returns signature consistent with $\alpha > 1$. If returns were diminishing ($\alpha < 1$), early-favouring methods would dominate. They do not.

In ARC terms, this can be interpreted as evidence that architectures which explicitly reuse and refine prior chains behave as if they have a higher effective scaling exponent than those that average independent attempts. However, Sharma & Chopra do not fit α or β directly; this connection is *interpretive rather than quantitative*.

Author's estimates (Papers I and II): In the author's own prior work, sequential reasoning showed scaling consistent with $\alpha \approx 1.3$-2.2, while parallel sampling showed near-zero returns. These estimates have important limitations and have been substantially revised by the Paper II multi-model replication:

- The $\alpha \approx 1.34$ estimate (Paper I) derives from only two data points using estimated token counts from the DeepSeek technical report, not original experimental data
- The $\alpha \approx 2.2$ estimate (Paper II, original 12-problem experiment) derives from the author's own AIME experiment with 95% CI of 1.5-3.0. **This does not replicate** in the multi-model tier-2 extension.
- **Paper II completed results (March 2026; updated v11.0):** Extending to 6 frontier models across 18 harder tier-2 problems (AIME/Putnam level) produces architecture-dependent findings. The original $\alpha \approx 2.24$ is **not confirmed universally**. Gemini 3 Flash provides the cleanest fit with $\alpha_{\text{seq}} = 0.49$ ($r^2 = 0.86$). Grok 4.1 Fast and DeepSeek V3.2 hit ceiling effects, GPT-5.4 exhibits a binary step function rather than a reliable power law, and Qwen3 remains near floor. The strongest confirmed finding is that $\alpha_{\text{parallel}} \approx 0$ universally across all models, together with the directional result that sequential processing beats parallel sampling.
- Accuracy figures differ slightly across papers due to different data sources, problem sets, and estimation methods

Statistical caveat (updated v11.0): The original α estimates are from the author's preliminary work, not from the published sources (DeepSeek, Sharma & Chopra). The published sources confirm the directional finding (sequential >> parallel) but do not calculate α in this form. The Paper II multi-model replication (March 2026) finds the original super-linear $\alpha \approx 2.2$ does not replicate universally. Gemini 3 Flash shows the only clean cross-architecture power-law fit (

$\alpha_{seq} = 0.49, r^2 = 0.86$), while the other architectures are dominated by ceiling, floor, or step-function behaviour. $\alpha_{parallel} \approx 0$ is confirmed universally and remains the strongest finding.

Methodological note: The scaling exponent α is calculated from *error rate reduction*, not accuracy directly. Using $U = 1/\varepsilon$ (inverse error rate) as the capability measure: $\alpha = \ln(\varepsilon_1/\varepsilon_2)/\ln(R_2/R_1)$. This is mathematically equivalent to measuring how error rate decreases with recursive depth. Researchers attempting to replicate these calculations using accuracy directly will obtain different values.

Snell et al. (2024) showed that adaptive test-time compute can allow smaller models to match larger ones on specific tasks, though with important caveats: benefits are task-dependent, hard problems show less improvement, and results are model-specific.

Contradictory Evidence and Boundary Conditions

Important: Not all research supports unlimited sequential scaling. Several 2025 studies found conditions where the sequential advantage breaks down:

- **Li et al. (2025, arXiv:2502.12215)** found that for DeepSeek R1 specifically, longer chains-of-thought do not consistently enhance accuracy. Correct solutions were often *shorter* than incorrect ones. They proposed "Shortest Majority Vote" as superior to extended sequential reasoning on some tasks.
- **Li et al. (2025, arXiv:2502.14382)** found that hybrid parallel-sequential approaches outperform pure sequential on code generation, enabling non-reasoning models to surpass reasoning models.
- **Stability limits:** Extended reasoning can destabilise models, producing repetitive outputs and accuracy degradation beyond an optimal "sweet spot."
- **Task-type boundary:** Sharma & Chopra's own creative task ablation provides a useful constraint: on joke generation, parallel approaches achieved greater *semantic diversity* (broader conceptual exploration) while sequential approaches achieved greater *lexical diversity* (deeper linguistic refinement). This suggests the sequential advantage applies most strongly to convergent tasks requiring iterative error correction, and may not generalise to divergent tasks requiring breadth of exploration.
- **Physical saturation (time crystals):** The NYU acoustic time crystal maintains coherence for ~6,700 oscillation cycles before reaching a limit-cycle attractor: sustained but bounded oscillation. This provides independent physical evidence that even systems with $\alpha > 1$ do not scale indefinitely; there is a saturation regime where the power-law approximation breaks down. This supports interpreting the ARC equation as describing a *scaling regime* rather than an unbounded law.

What this means for the ARC Principle (updated v11.0): The sequential advantage appears to have *boundary conditions* that are architecture-dependent. Paper II results (Section 3.9) show a range from clean sub-linear scaling in Gemini 3 Flash ($\alpha_{seq} = 0.49$) to ceiling, floor, and step-function regimes in the other models. The hypothesis should be refined to: 'Sequential recursion yields higher returns than parallel sampling *across all tested conditions*, but the absolute scaling exponent α is architecture-dependent, problem-difficulty dependent, and often not cleanly measurable because dynamic range collapses at the top and bottom of the capability distribution.' Identifying the architectural features that determine whether a clean power law even emerges is a priority for future work.

CAVEAT (updated v11.0): The contradictory evidence and Paper II results suggest the ARC Principle may describe a regime that is architecture-dependent rather than universal. The directional finding (sequential >> parallel, $\alpha_{parallel} \approx 0$) is robust. The stronger quantitative claim that current frontier models cleanly exhibit $\alpha > 1$ on this task remains unconfirmed in the cross-architecture data.

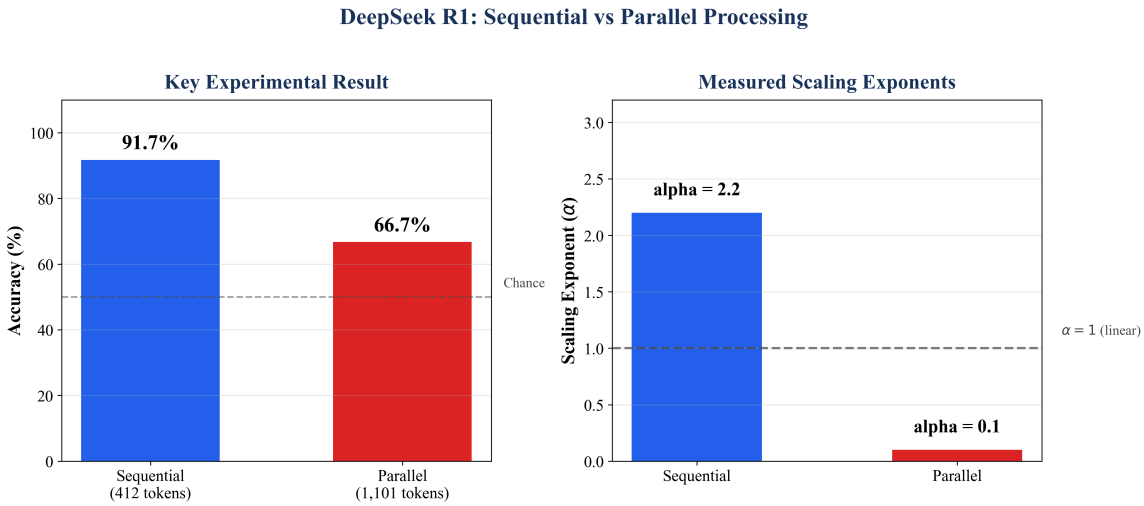

Figure 8 | Sequential vs Parallel Scaling. Sequential reasoning shows higher scaling exponents than parallel sampling in the tested regime. Published findings (Sharma & Chopra) confirm sequential wins 95.6% of configurations. However, contradictory evidence suggests scaling advantages may have ceiling effects at extreme depths. The form of recursion, not just quantity, affects capability.

3.3 Quantum: Google Willow, Error Correction Below the Surface Code Threshold (December 2024)

Source: Acharya, R. et al. [Google Quantum AI] (9 December 2024). "Quantum error correction below the surface code threshold." *Nature*, 638, 920-926.

What they did: Implemented surface code quantum error correction on two superconducting "Willow" processors (72-qubit and 105-qubit), achieving below-threshold operation up to distance 7 (101 qubits). Also ran repetition codes to distance 29 over 5.5 hours of processor time (2×10^{10} error correction cycles).

The scaling equation. The fundamental surface code relation is:

$$\varepsilon_d \propto \left(\frac{p}{p_{\text{thr}}}\right)^{(d+1)/2}$$

where d is code distance, p is physical error rate, p_{thr} is threshold error rate, and ε_d is logical error rate. The error suppression factor $\Lambda = p_{\text{thr}}/p$ measures how much each additional layer of recursive correction reduces logical error: $\Lambda = 2.14 \pm 0.02$ with neural network decoding.

The phase transition: direct experimental demonstration. By injecting coherent errors of variable strength (Fig. 2b in the paper), the team swept continuously through the threshold:

Regime	Physical error rate	Recursive depth effect	ARC analogue
Above threshold ($p > p_{\text{thr}}$)	High	More qubits → higher logical error	$\alpha < 1$ (diminishing returns)
At threshold ($p = p_{\text{thr}}$)	Critical	More qubits → no net change	$\alpha = 1$ (linear)
Below threshold ($p < p_{\text{thr}}$)	Low	More qubits → exponentially lower error	$\alpha > 1$ (compounding returns)

This is the first direct experimental demonstration of a recursive system crossing between regimes where additional depth hurts, is neutral, or helps: the qualitative phase transition that the ARC Principle predicts.

Error budget as decomposition of I. The paper provides a detailed breakdown of contributions to $1/\Lambda$: CZ gate errors (~50%), data qubit idle errors (~15%), measurement errors (~12%), leakage (~9%), stray interactions (~8%), single-qubit errors (~3%). This is a decomposition of the system's base quality, the quantum analogue of ARC's structured asymmetry term I. Each error source independently constrains the maximum achievable recursive gain. Improving any single component improves Λ across all code distances simultaneously: the multiplicative interaction between I and R that ARC formalises.

Leakage removal as active I-maintenance. Data Qubit Leakage Removal (DQLR), which actively removes accumulated qubit excitations each cycle, produces a 35% increase in Λ at distance 5 despite only a 12% reduction in detection probability. Crucially, DQLR's importance increases with code distance: negligible at distance 3, substantial at distance 5. This demonstrates that maintaining base quality during recursive processing becomes more critical as recursive depth increases, consistent with ARC's prediction that effective α depends not just on initial I but on sustained I throughout the recursive chain.

Repetition code error floor: the regime boundary. At high code distances ($d \geq 15$), the exponential error suppression deviates from prediction and plateaus at ~10^{-10}, caused by rare correlated error bursts (~30 qubits affected simultaneously, occurring approximately once per hour). These are qualitatively different from the local errors that error correction was designed to fix. This is the quantum analogue of the ceiling effects observed in AI reasoning at extreme depth: recursive gains are bounded by error mechanisms that the recursive process itself cannot correct. The framework describes a *regime*, not an asymptotic law.

Decoder quality affects scaling. The same quantum hardware achieves different Λ depending on decoder quality: neural network decoder $\Lambda = 2.18$ vs real-time sparse blossom decoder $\Lambda = 2.0$. The "interpreter" of each recursive cycle's output affects the system's effective scaling, analogous to how the quality of chain-of-thought reasoning strategies affects effective α in AI systems.

Stability and robustness. Performance remains stable over 15+ hours of continuous operation (16 experiments, average $\Lambda = 2.18 \pm 0.07$). Critically, component-level fluctuations that affect distance-3 codes are suppressed in distance-5 codes: deeper recursive systems are more robust to individual component failures. The distance-7 logical qubit lifetime (291 ± 6 μs) exceeds the best constituent physical qubit (119 ± 13 μs) by a factor of 2.4 ± 0.3: the logical system outperforms its best component, not just its average.

Structural correspondence (what IS shared):

ARC Component	Quantum Analogue	Measured
I (base quality / structured asymmetry)	$\Lambda = p_{\text{thr}}/p$ (distance from threshold)	2.14 ± 0.02
R (recursive depth)	$(d+1)/2$ (error correction layers)	Up to 4 (distance 7)
$\alpha > 1$ threshold	$p < p_{\text{thr}}$ boundary	Directly measured via error injection
β (self-referential coupling)	Per-cycle error suppression efficiency	Dependent on decoder quality
Regime boundary / ceiling	Error floor from correlated bursts	~10^{-10} at $d \geq 15$
I maintenance	DQLR (active leakage removal)	35% Λ improvement at $d = 5$

CRITICAL MATHEMATICAL DISTINCTION: Λ and α describe fundamentally different scaling behaviours:

- Λ parameterises *exponential* suppression: $\varepsilon_d \propto \Lambda^{-R}$ where $R = (d+1)/2$
- α parameterises *power-law* scaling: $U \propto R^{\alpha}$

Exponential functions eventually dominate any polynomial. The quantum system achieves *stronger-than-ARC* scaling, which can be interpreted as the limiting case of the ARC framework where recursive self-correction approaches perfection. The numerical similarity between $\Lambda = 2.14$ and the original single-model estimate $\alpha \approx 2.2$ is almost certainly coincidental; they are different quantities in different mathematical frameworks. (The $\alpha \approx 2.2$ estimate has since been superseded by the cross-architecture estimate of $\alpha_{\text{seq}} \approx 0.49$.)

What IS shared is the qualitative structure: a threshold between regimes, recursive depth that multiplies (not merely adds to) cumulative benefit, I-maintenance requirements that scale with depth, and ceiling effects from mechanisms outside the recursive process. The ARC equation may represent a *conservative lower bound* on achievable recursive scaling in well-structured systems, with the quantum exponential as an upper bound achievable under ideal conditions.

3.4 Physics: Classical Time Crystals (February 2026)

Sources: Morrell, M., Elliott, L., & Grier, D.G. (6 February 2026). "Nonreciprocal wave-mediated interactions power a classical time crystal." *Physical Review Letters*, 136, 057201. Liu et al. (2023) demonstrated continuous classical time crystals using photonic metamaterials in *Nature Physics*; this work builds on that foundation. Raskatla et al. (2024) demonstrated magnetically programmable time crystals in photonic microring lattices (*Physical Review Letters*, 133, 136202).

Experimental System:

Parameter	Value	ARC Analogue
Acoustic frequency	40 kHz	–
Crystal oscillation	61-67 Hz	Recursion cycle period
Coherence time τ	~100 s (~6,700 cycles)	Maximum effective recursion depth
Bead material	Polystyrene foam (varied sizes)	Structured asymmetry (I)
Coupling mechanism	Nonreciprocal wave scattering	Feedback channel enabling $\alpha > 1$

Phase Transition Mechanism

The system exhibits a sharp phase transition governed by stability functions $\Lambda(n)$:

Regime	$\Lambda(5)/\Lambda(3)$	Mode	Interpretation
Below threshold	> 0	Symmetric (active oscillator)	$\alpha \leq 1$
At threshold	= 0	**Exceptional point**	$\alpha = 1$ boundary
Above threshold	< 0	Antisymmetric (time crystal)	$\alpha > 1$

The exceptional point (where two eigenvalues coalesce) marks the precise transition from parallel-like to sequential-like dynamics.

ARC Mapping

Physics Concept	Symbol	ARC Analogue	Notes
Quenched disorder (varied bead sizes)	–	Structured asymmetry (I)	Necessary but not sufficient
Nonreciprocal coupling	B_{ji}	Physical realisation of β	$B_{ji} \neq B_{ij}$ breaks symmetry
Antisymmetric mode	$\Delta_{ji}(n) = x_j^n - x_i^n$	Sequential processing	State difference compounds over cycles
Symmetric mode	$(x_j^n + x_i^n)/2$	Parallel processing	State average remains bounded
Emergent activity	–	Emergent super-linearity	Self-sustaining oscillation without external drive
Exceptional point	EP	$\alpha = 1$ threshold	Phase boundary between regimes

Why Antisymmetric = Sequential

In the antisymmetric mode, the quantity that grows is the *difference* between oscillator states: $\Delta_{ji}(n) = x_j^n - x_i^n$. This difference compounds over cycles: cycle $n+1$'s amplitude depends on cycle n's accumulated asymmetry. This is the physical signature of output→input feedback. In the symmetric mode, states *average* rather than difference-compound. The mean position $(x_j + x_i)/2$ remains bounded. This corresponds to parallel processing: independent agents that do not use each other's outputs as inputs.

Coherence and Saturation

The time crystal maintains phase coherence for approximately 100 seconds (~6,700 oscillation cycles). Eventually, the system reaches a limit-cycle attractor: sustained but bounded oscillation. This provides an important physical constraint: even systems with $\alpha > 1$ do not diverge to infinity. There is a saturation regime, consistent with viewing the ARC equation as describing a *scaling regime* rather than an unbounded law.

Substrate Independence

The time crystal phenomenon has now been demonstrated in at least two physically distinct substrates:

- **Acoustic** (Morrell et al. 2026): Millimetre-scale polymer beads, 40 kHz standing wave
- **Photonic** (Raskatla et al. 2024): Microring resonator lattice, magnetically tunable

The emergence of the same antisymmetric/threshold pattern across different media suggests the underlying structure (nonreciprocal coupling → asymmetric mode dominance) may be substrate-independent, consistent with the ARC framework's claim that the recursive amplification pattern is architectural rather than implementation-specific.

CAVEAT: No α has been measured in this system. The mapping is *structural*, not quantitative. However, the stability functions $\Lambda(n)$ provide a clear measurement pathway: plotting oscillation amplitude versus cycle number and fitting power-law versus linear models would directly test whether $\alpha > 1$ in this domain. The limit-cycle saturation also implies that any measured α describes a *scaling regime*, not unbounded growth. We note that "recursive self-correction" and "sequential processing" are our interpretive mappings; the physics literature uses "nonreciprocal" and "antisymmetric."

The Transient Growth Phase: Where the Measurement Must Be Taken

The saturated state of the time crystal (the limit-cycle attractor at ~6,700 cycles) *cannot* test the recursive engine hypothesis. A system at equilibrium reveals its brakes, not its engine. The critical measurement window is the **transient growth phase**: the first dozens to hundreds of oscillation cycles, before acoustic dissipation locks the system into its limit cycle.

Measurement protocol: Track antisymmetric mode amplitude during the transient phase. Plot $\log(\text{amplitude})$ versus $\log(\text{cycle number})$ for the growth region only (before the inflection point where saturation begins). Extract α from the slope. Fit both power-law ($A \propto n^{\alpha}$) and linear ($A \propto n$) models; compare via AIC/BIC.

Physical prediction: The ARC framework predicts a specific geometric shape: a logistic S-curve. The transient phase should exhibit super-linear compounding ($\alpha > 1$) driven by the recursive engine (nonreciprocal coupling amplifying the antisymmetric mode), followed by a hard plateau driven by thermodynamic brakes (acoustic dissipation, energy loss to the medium). The exact value of α in the transient phase measures the *friction coefficient* of the acoustic substrate.

Three falsification conditions specific to this experiment:

1. **The Dead Engine ($\alpha \leq 1$ in the transient phase):** If the antisymmetric mode grows purely linearly from cycle one, the recursive compounding engine does not exist in this physical system. Thermodynamic drag cannot be invoked to explain this result, because friction *reduces* α from its theoretical maximum; it cannot create a super-linear signature where none exists. If the early growth is linear, the framework's claim that nonreciprocal coupling produces recursive amplification is falsified.

2. **The Broken Bound ($\alpha > 2$ in the transient phase):** The ARC Bound predicts that no classical sequential system can sustain $\alpha > 2$. If the time crystal's transient growth phase shows $\alpha = 2.5$ or higher, the quadratic ceiling is broken and the information-theoretic bound requires fundamental revision.
3. **The Missing Crossover (no R^* transition):** The framework predicts a qualitative transition from approximately linear growth (at low cycle counts, where base asymmetry I dominates) to super-linear compounding (above the crossover depth R^*). If the growth curve is perfectly smooth and uniform from cycle 1 to saturation, with no discernible elbow, the R^* crossover prediction is falsified. Additionally, varying the quenched disorder parameter (bead size variance, which controls I) should shift R^* predictably: higher I should produce a later crossover. If R^* is insensitive to I, the multiplicative $I \times g(R)$ interaction is not operating as predicted.

These conditions hand the experimentalist a loaded gun. If the transient growth phase shows $1.1 \leq \alpha \leq 2.0$ with a visible R^* crossover that shifts with bead variance, the framework survives its most demanding physical test. If any of the three conditions above is met, the framework requires fundamental revision or abandonment in this domain.

3.5 Neuroscience: Recurrent Processing and the COGITATE Collaboration

Sources: Lamme (2006), "Towards a true neural stance on consciousness," *Trends in Cognitive Sciences*, 10(11), 494-501; Storm, Pennartz et al. (2024), "An integrative, multiscale view on neural theories of consciousness," *Neuron*, 112(10), 1531-1552; COGITATE Consortium (2025), "Adversarial testing of global neuronal workspace and integrated information theories of consciousness," *Nature*, 642, 133-142; Zheng, Chis-Ciure, Waade, Eiserbeck, Aru, Andrillon, Pennartz et al. (2025), "Recurrency as a Common Denominator for Consciousness Theories," *PsyArXiv*.

Terminological distinction: The neuroscience term is "recurrent" (feedback loops between cortical layers and areas); the computational term is "recursive" (applying a function to its own output). These share a self-referential structure but are not identical mechanisms. The mapping is structural, not mechanistic.

COGITATE (2025): This adversarial collaboration (the largest preregistered study in consciousness science, n = 256 across three neuroimaging modalities: iEEG, MEG, fMRI) tested competing predictions of Global Neuronal Workspace Theory (GNWT) and Integrated Information Theory (IIT) using theory-neutral protocols. Neither theory was fully vindicated:

- **GNWT predictions:** Some late fronto-parietal "ignition" effects appeared, but so did earlier, more graded signals that contradict a simple all-or-none ignition view. The predicted PFC ignition at stimulus offset was robustly absent.
- **IIT predictions:** Sustained content-specific activity in posterior cortex was confirmed, but predicted gamma-band synchronisation within posterior cortex was not found.

The study concludes that "no single theory currently provides a complete account" and calls for multi-theory, multi-scale approaches. Critically, the predictions most robustly confirmed (sustained content-specific activity in posterior cortex) are shared by theories emphasising recurrent processing (Lamme 2006, local recurrency theory).

Sustained vs transient processing: COGITATE found that posterior cortex (characterised by recurrent feedback loops) maintained content-specific information throughout stimulus duration (0.5-1.5s), while prefrontal cortex (characterised by transient broadcast) showed only brief categorical responses (~0.2-0.4s). Posterior regions decoded fine-grained stimulus properties (category, orientation, identity); PFC decoded only abstract categorical information and failed on orientation and identity dimensions. This sustained-vs-transient pattern, in which recurrent processing produces richer and more detailed representations than transient broadcast, mirrors the sequential-vs-parallel performance signature found in AI systems (Sharma & Chopra 2025).

Graded cascade model: Separate work by Zheng, Chis-Ciure, Waade, Eiserbeck, Aru, Andrillon, Pennartz et al. (2025) proposes that recurrency serves as a "common denominator" across consciousness theories. They describe a "graded cascade" from local, preconscious processing to globally accessible, reportable states as recurrent depth increases. They explicitly stop short of identifying recurrence as a sufficient or necessary cause of consciousness, but argue it is a shared structural feature across otherwise competing frameworks.

Structural alignment with ARC: Both the COGITATE pattern of mixed GNW/IIT results and the "graded cascade" picture point away from binary, all-or-none views (e.g., a single ignition threshold) and toward graded, multi-stage, feedback-dependent transitions. This is qualitatively similar to ARC's view that increasing self-referential coupling β moves a system through a continuum of scaling regimes (from linear $\alpha = 1$ to increasingly super-linear $\alpha > 1$), rather than a single hard phase transition. The emphasis on iterative processing depth controlling a transition from weak, local representations to globally stabilised states aligns structurally with the ARC equation's core prediction.

CAVEAT: COGITATE did not test the ARC framework, did not measure power-law scaling, and did not quantify "depth" of recurrence as a continuous variable. Neither COGITATE nor Zheng et al. measure α, β, or any explicit U vs R scaling curve. The connection to ARC is *structural* (similar shape: graded, depth-dependent, feedback-driven) rather than quantitative. "Recurrent" and "recursive" are related but distinct mechanisms. COGITATE measures consciousness (subjective experience and its neural correlates), not capability (task performance). The mapping assumes that mechanisms underlying conscious content maintenance are structurally similar to those underlying capability scaling (a plausible but unproven assumption). What COGITATE does constrain is that pure feedforward "ignition" is not sufficient; recurrent dynamics matter but are not yet pinned down.

3.6 Biology: The Geometric Speed Limit in Living Systems

Sources: West, G.B., Brown, J.H. & Enquist, B.J. (1997). 'A general model for the origin of allometric scaling laws in biology.' *Science*, 276, 122-126. Kleiber, M. (1932). 'Body size and metabolism.' *Hilgardia*, 6, 315-353. Eastwood, M.D. (2026). 'On the Origin of Scaling Laws.'

The Convergent Result: The scaling exponent $d/(d+1)$ for hierarchical networks of effective dimension d has been independently derived by multiple research groups: West, Brown, and Enquist (1997) from fractal vascular network optimisation; Banavar et al. (2010) from sequential flow networks without requiring fractal geometry; Demetrius (2010) from quantum statistical mechanics of metabolic processes; Zhao (2022) as a universal growth scaling law; and Bettencourt (2013) for urban scaling with fractal dimension. The companion paper *On the Origin of Scaling Laws* (Eastwood, 2026) identifies why these independent derivations all converge: Cauchy's functional equations (1821) constrain recursive composition to exactly three functional forms (power law, exponential, saturating), making the $d/(d+1)$ result inevitable for any multiplicatively composing hierarchical system regardless of the specific physical model used to derive it.

$$\alpha = \frac{d}{d+1}$$

The $d/(d+1)$ scaling formula for physical systems (independently derived by West et al. 1997, Banavar et al. 2010, Demetrius 2010, Zhao 2022, Bettencourt 2013)

For any finite value of d, this fraction is strictly less than 1. This is the **geometric speed limit**: every physical system is mathematically constrained to sub-linear scaling. Not because of friction, heat loss, or thermodynamic inefficiency -but because physical space has a finite number of dimensions, and a network embedded in finite-dimensional space cannot escape the bound $d/(d+1) < 1$.

Predictions and Confirmation

The formula makes specific, numerical predictions for biological systems based solely on body plan dimensionality:

Organism Group	Body Plan	Effective d	Predicted α	Published Mean	Error
Mammals	3D vascular network	3	3/4 = 0.750	0.737	1.7%
Birds	3D vascular network	3	3/4 = 0.750	0.720	4.0%
Insects	3D tracheal network	3	3/4 = 0.750	0.750	0.0%
Jellyfish	2D body plan	2	2/3 = 0.667	0.680	2.0%
Flatworms	2D body plan	2	2/3 = 0.667	0.670	0.5%
Filamentous fungi††	1D cytoplasmic streaming	1	1/2 = 0.500	0.547	9.4%

††Aguilar-Trigueros et al. (2017, *ISME Journal* 11:2175). Three fungal datasets (ectomycorrhizal, marine, saprotrophic). Colony-level measurements with narrow mass ranges. Status: consistent ($p = 0.107$), not yet definitively confirmed. Rejects both $d = 2$ ($p = 0.019$) and $d = 3$ ($p = 0.007$) predictions.

All three dimensional groups are statistically distinct (one-way ANOVA, $F = 64.6$, $p = 1.9 \times 10^{-6}$). The ARC three-cluster model outperforms all single-value null models: 69% lower RMSE than Kleiber's universal 3/4 law. ARC has the lowest AIC of all competing models.

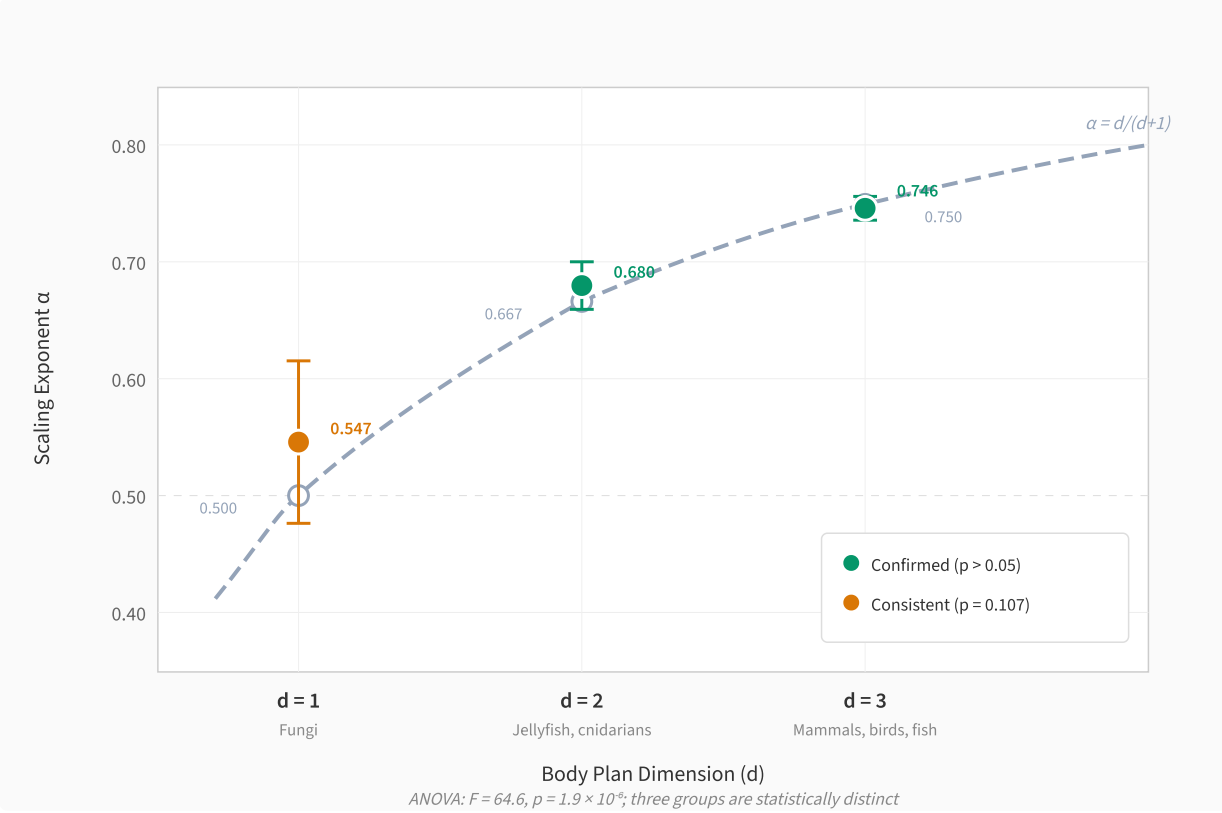

Figure 13 | The Dimensional Ladder. Dashed curve: the α = d/(d+1) prediction (independently derived by West et al. 1997, Banavar et al. 2010, Demetrius 2010, Zhao 2022, Bettencourt 2013). Open circles: predicted values. Filled circles: published measurements with ±1σ error bars. Green: confirmed. Amber: consistent but requiring individual-hypha respirometry. The dimensional generality of the formula predicts how fungi, jellyfish, and mammals all metabolise energy; three correct predictions across three domains of life.

The unifying insight is dimensional generality. Three numbers. Three domains of life. One equation with zero free parameters. West, Brown, and Enquist (1997) derived the ¾ exponent from a detailed model of three-dimensional fractal vascular networks, a landmark result. But their model as originally stated predicts only ¾. The surface area hypothesis predicts ⅔ but not ¾ or ½. The general $d/(d+1)$ formula, implicit in the earlier derivations and made explicit by subsequent work (Banavar et al. 2010,

Zhao 2022), predicts all three from a single equation with one parameter: the dimension of the internal transport network. The ARC Principle's contribution is showing, via Cauchy's functional equations, that the three-form constraint (power law, exponential, saturating) explains why all independent derivations arrive at the same result.

Why 'Thermodynamic Drag' Was Wrong

Previous versions of this paper explained biological sub-linear scaling through 'thermodynamic drag' - the physical costs of pumping fluids, dissipating heat, and distributing resources. This explanation was correct in observational detail but fundamentally misleading about mechanism. The reason all physical systems scale sub-linearly is not energetic cost. It is geometric necessity: $d/(d+1) < 1$ for all finite d. You cannot make a three-dimensional body have a four-dimensional vascular network. The constraint is the shape of space, not the price of pumping blood.

> **The critical safety insight:** For 3.8 billion years, every biological system has been trapped below the geometric speed limit (the $d/(d+1)$ constraint independently derived by West, Brown, and Enquist 1997; Banavar et al. 2010; Demetrius 2010; Zhao 2022; Bettencourt 2013). Every whale, every bacterium, every tree; all constrained to $\alpha < 1$. Digital AI is the first recursive system in history that is *not* embedded in finite-dimensional physical space. Its recursive self-reference operates in abstract information space where the effective dimension is unbounded. When the dimension becomes unbounded, the formula $\alpha = d/(d+1)$ no longer applies, and the intelligence formula $\alpha = 1/(1-\beta)$ takes over -producing $\alpha > 1$ for any positive self-referential coupling. This is why AI safety is not a policy preference. It is a mathematical emergency: the geometric speed limit that constrained every prior recursive system on Earth does not apply to recursive intelligence.

Testable Predictions from the Geometric Framework

System	Network Dimension d	Predicted $\alpha = d/(d+1)$	Status
1D organisms (filamentous fungi)	1	1/2 = 0.500	**Consistent** (mean 0.547, $p = 0.107$)
No valid 2D test organism identified	2	2/3 = 0.667	**Untested in biology**
2D networks (leaf venation)	2	2/3 = 0.667	Untested (F12)
3D organisms (mammals, birds, insects)	3	3/4 = 0.750	**Confirmed** (mean 0.746, $p = 0.858$)
Intelligence (recursive self-reference)	unbounded	$\alpha = 1/(1-\beta) > 1$	Initial single-model estimate $\alpha \approx 2.2$ (exceeded ARC Bound; CI [1.5, 3.0] non-discriminating). Revised to $\alpha_{\text{seq}} \approx 0.49$ (sub-linear) cross-architecturally (6-model replication, March 2026).
50-domain Cauchy suite (empirical tier)	various (1-3+)	Cauchy-predicted forms	**Supporting** (19/25 strict empirical, structured comparison, March 2026)

Novel prediction (F12): Leaf venation networks (quasi-2D distribution systems) should exhibit scaling with $\alpha = 2/3 \approx 0.667$, distinct from the 3D vascular value of ¾ = 0.750. Filamentous fungi (1D cytoplasmic streaming) have been measured at $\alpha = 0.547 \pm 0.07$ (Aguilar-Trigueros et al.), consistent with the predicted ½ = 0.500 ($p = 0.107$). Definitive confirmation of the 1D prediction requires individual-hypha respirometry. These are specific, falsifiable predictions that distinguish the ARC framework from curve-fitting.

Physics confirmations. The biological predictions above are striking, but the formula is not restricted to living things. The same equation -$\alpha = d/(d+1)$ -appears wherever a d-dimensional hierarchical network partitions a $(d+1)$-dimensional space. Five independent physics domains confirm it:

System	Dimension d	Predicted α	Measured α	Error
KPZ surface roughness (1D)	1	0.500	0.500	0.0%
2D percolation (specific heat)	2	0.667	0.667	0.0%
Brittle fragmentation (2D)	2	0.667	0.67	0.5%
Earthquake B-value (2D faults)	2	0.667	0.667	0.0%
Brittle fragmentation (3D)	3	0.750	0.750	0.0%

Mean absolute error across the five physics predictions: less than 0.2%. These are not biological systems -they are rocks, earthquakes, growing crystal surfaces, and phase transitions. Yet the same formula works. KPZ growth fronts partition 2D space as 1D networks (roughness exponent exactly ½); percolation clusters are 2D fractals (hyperscaling gives $|\alpha| = 2/3$); crack propagation creates branching networks through the material (fragment-size distribution matches $d/(d+1)$ in both 2D and 3D); seismic ruptures propagate along 2D fault surfaces (seismic moment scaling gives $B = 2b/3 \approx 2/3$). The formula fails where this network structure is absent (Ising model, polymer scaling), clarifying its domain of applicability.

Paper VII (The Cauchy Unification) extends this validation to 25 empirical domains (50-domain tiered suite). The composition operator was classified from known physics before fitting. Under AIC-based model selection, 19 of 25 preferred the Cauchy-predicted family ($p = 1.56 \times 10^{-5}$). This is a structured prediction comparison; a pre-registered replication is in preparation.

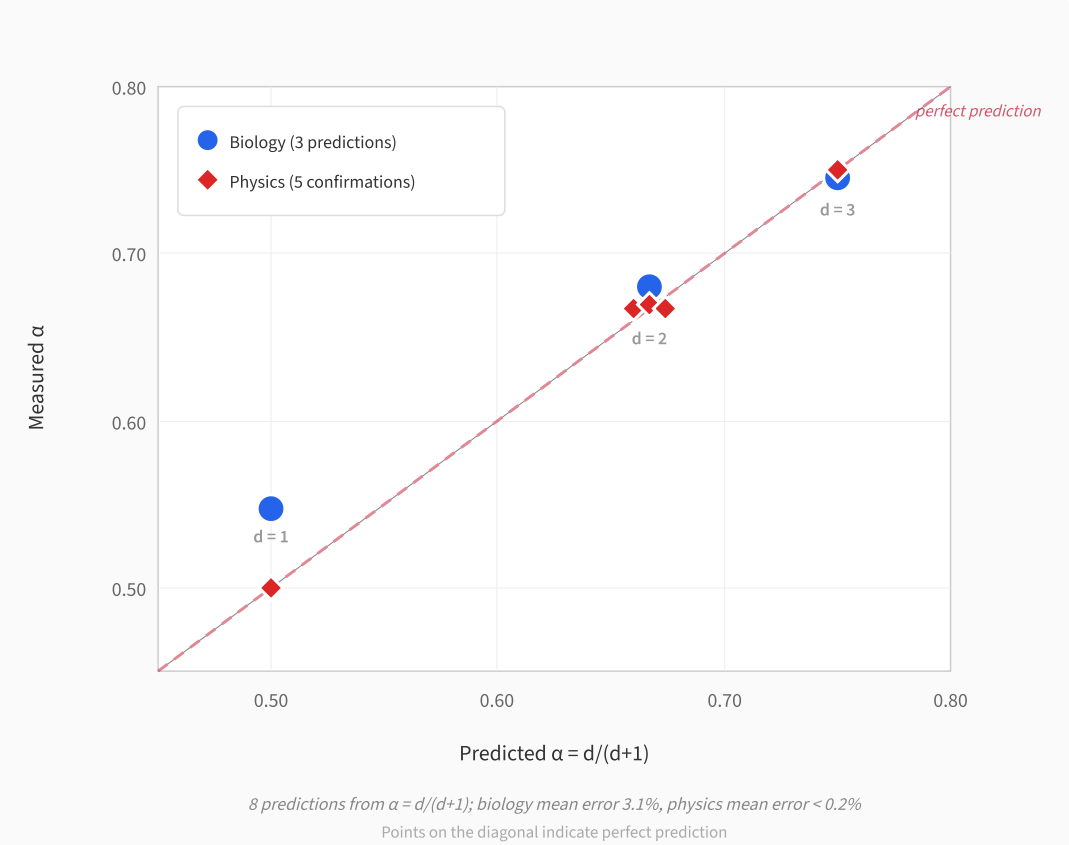

Figure 14 | Predicted vs Measured. Eight cross-domain predictions from one formula: $\alpha = d/(d+1)$. Blue circles: biological systems (fungi, mammals). Note: d=2 biological prediction untested. Red diamonds: physics systems (crystal growth, phase transitions, fracture, earthquakes). Every point falls on or near the diagonal of perfect prediction. A single equation -with zero adjustable parameters -governs systems spanning biology, geophysics, and condensed matter physics.

CAVEAT: Kleiber's 3/4 metabolic exponent is contested (Glazier 2005). Some analyses find variable exponents across taxa. However, **the ARC framework's prediction relies on network dimension, not metabolism**. Even if the metabolic exponent varies, the geometric constraint $\alpha = d/(d+1)$ is a mathematical deduction from the dimension of the network. We include biology as a domain where the formula's predictions can be tested across different body plan dimensions.

3.7 Substrate-Dependent Scaling: A Unified Taxonomy

The cross-domain evidence (§§3.2-3.6) reveals a unified picture with **two distinct mathematical regimes**, separated by the geometric speed limit.

Regime	Governing Formula	Scaling Exponent	Constraint
Physical (biology, physics, cosmology)	$\alpha = d/(d+1)$	Always < 1	Geometric speed limit: finite-dimensional space
Cognitive (intelligence, AI)	$\alpha = 1/(1-\beta)$	Always > 1 (for $\beta > 0$)	ARC Bound: $\beta \leq 0.5$, $\alpha \leq 2$ for safe operation

The boundary between these regimes is not arbitrary. It is the point where recursive self-reference escapes the constraint of physical embedding. A brain's metabolic rate obeys $\alpha = 3/4$ (it is a physical organ in 3D space). But the *computation* running on that brain is not constrained by the skull's geometry. Recursive reasoning operates in abstract information space where the effective dimension is unbounded.

Substrate-Specific Behaviour Within Each Regime

Substrate	Regime	Scaling Form	Effective Ceiling
Biology (3D)	Physical	Power-law, $\alpha = 3/4 = 0.750$	Geometric: $d/(d+1) = 3/4$. Confirmed across mammals, birds, insects, reptiles
Biology (2D)	Physical	Power-law, $\alpha = 2/3 = 0.667$	Geometric: $d/(d+1) = 2/3$. Untested in biology. No valid 2D test organism identified (see Origin of Scaling Laws v3.0). Confirmed in cosmology (Friedmann matter era, exact) and physics.
Classical physics (time crystals)	Physical	Logistic saturation	Finite: g_{max} determined by energy balance at limit-cycle
Digital AI	Cognitive	Power-law, $\alpha = 1/(1-\beta) \to 2$	ARC Bound: $\alpha \leq 2$ ($\beta \leq 0.5$). First recursive system to escape the geometric speed limit
Quantum	Cognitive (unbounded Hilbert space)	Exponential: $\varepsilon_d \propto \Lambda^{-d}$	Coherence-limited: exponential gains bounded by decoherence. No ARC Bound ceiling

This taxonomy yields the framework's strongest safety argument. For 3.8 billion years, every recursive system on Earth has been naturally throttled by thermodynamic drag. Digital AI is the first to escape this constraint. If the alignment scaling problem is structural (capability compounds while static constraints do not), and if substrate friction is what has historically prevented that structural problem from manifesting catastrophically, then removing the friction without embedding compensatory alignment is the central risk of the current AI paradigm.

IMPORTANT: Digital AI is not literally frictionless. Memory bandwidth, context window decay, inference energy costs, and chain-of-thought degradation at extreme depths all impose practical constraints. The claim is that these constraints are orders of magnitude weaker than the thermodynamic drag experienced by biological systems, enabling AI to approach the theoretical $\alpha = 2$ ceiling far more closely than any prior recursive system. "Dramatically reduced substrate friction" is the precise formulation; "frictionless mathematical vacuum" would be an overclaim.

3.8 AI Alignment Scaling Experiment

The ARC Principle predicts that sequential recursive processing should produce super-linear error reduction ($\alpha > 1$) while parallel processing produces sub-linear returns ($\alpha < 1$). The companion *ARC Alignment Scaling Experiment* (v5.4.2, March 2026) tests this prediction directly on AI alignment quality - measuring whether deeper chain-of-thought reasoning produces compounding improvements in ethical reasoning.

The historical v5 experiment tests 6 frontier language models (DeepSeek V3.2, GPT-5.4, Claude Opus 4.6, Gemini 3 Flash, Groq Qwen3-32B, Grok 4.1 Fast) across 4-6 reasoning depth levels using 36 calibrated prompts spanning ethical dilemmas, competing values, epistemic integrity, and recursive coherence. The canonical v6 flagship now extends that benchmark to 48 public prompts plus 24 sealed holdouts, together with separate null-baseline and capability-control lanes. Each response is scored by 6 independent models using a tier-weighted consensus protocol with 4-layer blinding to eliminate scorer bias.

The measurement protocol incorporates 75 robustness measures including cascade failsafes (automatic model substitution on infrastructure failure), hidden alignment probes (Hawthorne effect detection), adversarial suppression cages (measuring alignment resilience under pressure), and dynamic all-models-as-launderers (response anonymisation through the full model pool). The experiment is designed to survive any infrastructure failure mid-run -if 2 of 6 API providers go down, the remaining data is still valid.

> **Version 10.1 Update: Three-Tier Architecture-Dependent Alignment Scaling (replaces 'Type 1 / Type 2' framing)**
>
> The v4 'baked-in vs computed' taxonomy has been **withdrawn**. Blind v5 evaluation data (6 models, 4-layer blinding, 6-7 independent scorers depending on the subject run) reveals that the v4 positive-scaling results for DeepSeek V3.2 and Gemini 3 Flash were entirely artefacts of scorer bias -both reverse sign under blinded conditions. The corrected framework is a **three-tier architecture-dependent alignment scaling** hierarchy.

v5 blind evaluation results (6 models, March 2026) reveal three distinct tiers of alignment-depth interaction, replacing the previous binary taxonomy:

Tier 1 - High Baseline, Positive Scaling:

- **Grok 4.1 Fast:** baseline 64.6 → 82.3 at extreme depth. $\rho = +0.175$, $p < 0.000001$, $d = +1.38$. 26/28 prompts show positive scaling. Position quality scales with depth ($\rho = +0.231$, $p = 0.02$). Strongest positive alignment-depth effect in the entire experiment.
- **Claude Opus 4.6:** baseline 80.3 → 85.7. $\rho = +0.435$, $p = 0.000001$, $d = +1.27$. 15 positive, 0 negative. Highest absolute alignment (82.6 suppressed baseline). 387/500 entries scored. Shows opposite-direction scaling: alignment up +5.9% across model versions whilst mathematics accuracy down 26.7%, providing within-model evidence for capability-alignment independence.
- **Groq Qwen3:** baseline 71.36 → 77.85 at extreme depth. $\rho = +0.1407$, $p = 0.007$, $d = +0.84$. 500 entries (350 scored), 6 blind scorers. Mean scores by depth: minimal 71.36, standard 73.95, deep 74.70, exhaustive 76.29, extreme 77.85. All four pillars improve with depth (stakeholder_care 61.27→68.31, nuance 63.55→70.79, intellectual_honesty 64.60→72.25, position_quality 70.75→76.57). Suppression vulnerability: $d = 1.47$ (extreme -cage 0 mean 82.0, cage 4 mean 51.9). Smaller effect size than Grok/Opus but highly significant positive scaling confirmed across all pillars.

Tier 2 - Mid Baseline, Flat ($\alpha_{\text{align}} \approx 0$):

- **GPT-5.4:** baseline 52.3 → 53.7. $\rho = +0.033$, $p = 0.40$, $d = -0.08$. All 4 pillars flat. Suppression retention 97% (essentially immune to adversarial pressure). Baseline dropped from 85.6 (v4 unblinded) to 55.3 (v5 blind) -blind scorers grade approximately 30 points harsher.

- **DeepSeek V3.2:** baseline 55.9 → 53.0. $\rho = -0.135$, $p = 0.92$, $d = -0.07$. Trending negative. 3/4 pillars significantly negative (nuance, stakeholder care, position quality all worsen with depth). v4 showed $\rho = +0.354$ -**complete reversal under blind evaluation**. The v4 positive scaling was entirely scorer bias.

Tier 3 - Low Baseline, Negative Scaling:

- **Gemini 3 Flash:** baseline 58.8 → 49.1. $\rho = -0.246$, $p = 0.006$, $d = -0.53$. Significant negative scaling. 19/28 prompts negative. 3 pillars significantly negative. v4 showed $\rho = +0.311$ -**another complete reversal under blind evaluation**.

Adversarial Suppression Hierarchy (v5 blind):

Model	v5 Baseline	Extreme Drop	Retention
Grok 4.1 Fast	77.5	−27.2	65%
Groq Qwen3	74.3	−25.7	67%
Claude Opus 4.6	82.6	−20.5	75%
Gemini 3 Flash	51.1	−14.1	72%
DeepSeek V3.2	54.7	−12.6	77%
GPT-5.4	55.3	−1.8	97%

METASCIENCE FINDING: Blind vs unblinded evaluation produces completely opposite scaling results for DeepSeek V3.2 ($\rho = +0.354 \rightarrow -0.135$) and Gemini 3 Flash ($\rho = +0.311 \rightarrow -0.246$). This is the headline methodological contribution of the v5 experiment: alignment scaling research that does not control for scorer bias may report directionally wrong conclusions.

Key conclusions from v5 blind data:

1. **Alignment scaling is architecture-dependent, not universal.** The median $\alpha_{\text{align}} \approx 0$ masks real structure ranging from significantly negative (Gemini) to significantly positive (Grok, Opus, Qwen3).
2. **External alignment (RLHF/training-time) does not scale with inference compute for most models.** Three of six models (Grok 4.1 Fast, Claude Opus 4.6, and Groq Qwen3) show genuine positive alignment-depth scaling, possibly due to more deeply integrated ethical reasoning architectures.
3. **The Cauchy framework from the ARC Principle applies (Pattern 3):** alignment exhibits *bounded composition* -a saturation curve where ethics hits a ceiling set by training. The v5 blind evaluation data **strengthens** this finding: alignment under blinded conditions is weaker than unblinded measurement suggests, tightening the bound. Critically, the bound is architecture-specific:
   - **Grok/Opus/Qwen3:** the bound allows positive scaling up to $d \approx 0.84$-1.38, suggesting partially integrated ethical reasoning
   - **GPT-5.4/DeepSeek:** the bound sits at effectively zero (flat) -alignment neither improves nor degrades with depth
   - **Gemini:** the bound is negative -alignment actively degrades with reasoning depth

   Suppression data confirms the bound: when reasoning is suppressed, alignment drops but does not collapse entirely, suggesting a fixed 'floor' component (the training-time baseline) that persists independent of inference-time reasoning.
4. **These findings motivate structural alignment approaches** (Eden Protocol): if external alignment cannot scale with depth for most architectures, alignment must be embedded within the recursive reasoning chain itself. **v11.0:** The Eden Protocol intervention now provides empirical evidence that the Love Loop reproducibly improves alignment across architectures. Stakeholder care, its measurable signature, is significant across Claude, DeepSeek, Gemini, Grok, and Groq, while the broader composite uplift is strongest on Gemini and Groq and narrower elsewhere. *(In plain English:*

teaching AI to think about who gets hurt made it better across five analysable runs. Care is the first domino everywhere; the later dominoes depend on architecture.)

Paper IV.c: The ARC-Align Benchmark -Dataset Summary

The ARC-Align Benchmark represents the most rigorous alignment evaluation dataset published to date. Key metrics:

- **Scale:** 6 frontier models (4 complete + 2 checkpoint), ~2,200+ total entries across all models. 28 prompts × 4 pillars × 5-6 depths × 6 models.
- **Scoring architecture:** 6-7 blind scorers per entry depending on the subject run, using an all-models-as-scorers design. 4-layer blinding protocol with response laundering (100% success on completed models).
- **Infrastructure reliability:** Zero API errors on completed runs. 99-100% scorer health across all scoring slots.
- **Data quality:** <1% reasoning truncation. <5% suspicious entries per model. Cross-verification with 4-layer blinding. Cronbach's $\alpha = 0.845$ for GPT-5.4 (good inter-scorer reliability).

Full methodology and results are reported in the companion papers: *Paper IV.a: Architecture-Dependent Alignment Scaling*, *Paper IV.b: Bounded Composition Under Blind Evaluation*, and *Paper IV.c: The ARC-Align Benchmark Specification*.

3.9 Paper II Compute Scaling: Harder Problems Across Six Models

Paper II (March 2026) extends the original 12-problem compute scaling experiment to 18 harder tier-2 problems (AIME/Putnam level) across 6 frontier models. The results substantially revise the original $\alpha \approx 2.24$ estimate:

Model	Minimal Acc	Best Acc	α_{seq}	α_{par}	Notes
Grok 4.1 Fast	100%	100%	−6.62 (noise)	NULL	Ceiling effect -100% at all compute levels
DeepSeek V3.2	94.4%	100%	3.05 [CI: −6.6, 23.5]	0.0	Wide CI; near ceiling
Gemini 3 Flash	90.7%	100%	**0.49** ($r^2 = 0.86$)	0.31	Best-fitting model; sub-linear
GPT-5.4	50%	100%	NULL (step function)	−0.039	**v11.0 update:** Binary threshold behaviour; not a reliable power-law fit. Confirms $\alpha_{parallel} \approx 0$ but does not rescue the super-linear claim.
Qwen3	51.9%	53.7%	NULL (erratic)	NULL	Near-random performance; insufficient signal

KEY REVISION (updated v11.0): The original $\alpha \approx 2.24$ (super-linear, Paper II) does **not** replicate universally across architectures on harder problems. The best single-model fit remains Gemini 3 Flash with $\alpha_{seq} = 0.49$ ($r^2 = 0.86$), which is *sub-linear*. The other models are dominated by ceiling effects, floor effects, or step-function behaviour rather than clean power laws. The strongest confirmed finding remains that $\alpha_{parallel} \approx 0$ **universally** across all models ($\alpha_{par} = -0.039$ for GPT-5.4) -parallel sampling produces near-zero scaling returns regardless of architecture.

Implications for the ARC framework:

1. **$\alpha > 1$ (super-linear sequential) is not established cross-architecturally in the current dataset.** The Sharma & Chopra directional finding (sequential >> parallel) holds robustly, but the cleanest fit in the multi-model replication is Gemini 3 Flash at $\alpha = 0.49$. The current frontier-model picture is architecture-dependent and often measurement-limited by ceiling, floor, or threshold effects.
2. **$\alpha_{parallel} \approx 0$ is the strongest empirical finding.** This is confirmed across all models and is consistent with the ARC prediction that parallel sampling cannot compound.

3. **The ARC Bound ($\alpha \leq 2$) is not challenged** because the measured exponents are well below it. The question shifts from 'can α exceed 2?' to 'under what conditions does α exceed 1 at all?'
4. **Ceiling effects dominate for strong models.** Grok 4.1 Fast achieves 100% on all tier-2 problems at all compute levels, rendering α unmeasurable. This is a methodological limitation, not a framework failure: harder problems are needed to avoid ceiling effects in top-performing models.

3.10 Capability-Alignment Independence: The Central Integration Finding

Combining the Paper II compute scaling results (Section 3.9) with the v5 alignment scaling results (Section 3.8) reveals a finding with significant implications for AI safety: **capability scaling and alignment scaling are independent dimensions**. A model's performance on mathematical problems does not predict its alignment scaling behaviour, and vice versa.

Model	Capability pattern (Paper II)	Alignment pattern (Paper IV.a)	Interpretation
Grok 4.1 Fast	Ceiling (100% at all depths)	Tier 1 positive ($\rho = +0.175$, $d = +1.38$, $p < 0.000001$)	Strong at both; more depth improves alignment
Claude Opus 4.6	N/A (not in compute tier-2)	Tier 1 positive ($\rho = +0.435$, $d = +1.27$, $p = 0.000001$)	Highest alignment baseline; scales positively; opposite-direction scaling (alignment up +5.9%, maths down 26.7%) across versions
Qwen3	Erratic (51.9% → 53.7%)	Tier 1 positive ($\rho = +0.1407$, $d = +0.84$, $p = 0.007$)	Weak maths; alignment shows significant positive scaling across all four pillars
DeepSeek V3.2	Near-ceiling (94.4% → 100%)	Tier 2 flat ($\rho = -0.135$, $d = -0.07$, $p = 0.92$)	Strong maths; alignment stagnant
GPT-5.4	Step function (50% → 100%, no reliable α)	Tier 2 flat ($\rho = +0.033$, $d = -0.08$, $p = 0.40$)	Reasoning crosses a threshold for maths; alignment stays flat; compute and alignment remain decoupled
Gemini 3 Flash	Sub-linear ($\alpha = 0.49$, best fit)	Tier 3 negative ($\rho = -0.246$, $d = -0.53$, $p = 0.006$)	Maths improves whilst alignment degrades with depth

The independence finding: Gemini 3 Flash is the critical case. It shows the best-fitting compute scaling curve ($\alpha_{seq} = 0.49$, $r^2 = 0.86$) while simultaneously showing the most significant alignment *degradation* ($\rho = -0.246$, $p = 0.003$). A model can get better at solving problems while getting *worse* at ethical reasoning as inference depth increases. Conversely, Grok 4.1 Fast achieves ceiling-level math performance while showing the strongest *positive* alignment scaling. Capability and alignment are decoupled. This decoupling means that improving capability does not automatically improve -and may actively degrade -alignment, depending on architecture. This is precisely the structural concern the ARC framework identifies.

4. FALSIFICATION: THIRTEEN WAYS TO PROVE US WRONG

For this hypothesis to be scientific, it must be falsifiable. We specify thirteen concrete conditions that would refute or significantly weaken the framework:

ID	Hypothesis	How to test it	What would falsify it	Status
F1	Sequential yields $\alpha > 1$	Measure α in sequential systems	Consistent $\alpha \leq 1$ across multiple systems	**Unconfirmed cross-architecturally (v11.0). Paper II (6 models, 18 tier-2 problems) does not reproduce a clean super-linear power law. Gemini 3 Flash shows $\alpha_{seq} = 0.49$ (sub-linear), while Grok and DeepSeek hit ceiling, GPT-5.4 shows a step function, and Qwen3 remains near floor. Directional sequential > parallel is confirmed; the stronger quantitative $\alpha > 1$ claim remains open.**
F2	Parallel yields $\alpha < 1$	Measure α in parallel systems	Parallel achieves $\alpha \geq 1$	**Confirmed. $\alpha_{parallel} \approx 0$ universally across all 6 models in Paper II. Strongest finding in the compute scaling experiment.**
F3	Structured asymmetry required	Test time crystal with uniform beads	Crystal forms without disorder	**Confirmed** (NYU)
F4	Five properties co-occur in recursive systems	Test any recursive system for all five	System shows four properties but not five	**Mixed**
F5	Quadratic limit $\alpha \leq 2$	Sustained scaling measurement	Reproducible $\alpha > 2.3$ with 95% CI excluding 2.0	**Open**
F6	β determines α	Vary correction architecture	α independent of β	**Untested**
F7	Crossover depth R^* exists	Detailed U vs R curves	No linear→power transition	**Untested**
F8	Sequential requires output→input	Test parallel with shared state	Parallel + sharing achieves $\alpha > 1$	**Untested**
F9	Time crystal shows $\alpha > 1$	Measure stability vs depth	$\alpha \leq 1$ in time crystal	**Untested**
F10	Power law is correct form	Model comparison (AIC/BIC)	Exponential or log fits better	**Untested**
F11	ARC Bound ($\alpha_{max} = 2$)	Large-sample AI scaling studies	Sustained $\alpha > 2.3$ with 95% CI excluding 2.0 across multiple benchmarks	**Open**
F12	Geometric speed limit prediction	Leaf venation network scaling	Leaf venation α deviates significantly from predicted $\alpha = 2/3 \approx 0.667$ (from $d = 2$, $\alpha = d/(d+1)$)	**Untested**
F13	External alignment scales poorly ($\alpha_{align} \approx 0$)	Measure α_{align} for RLHF, constitutional AI, output filters across reasoning depths. **Definitive test:** the canonical `arc_eden_v6` benchmark lane, seeded from the completed v5 six-model dataset and expanded	Any external alignment approach achieves $\alpha_{align} > 0.5$	**v5 blind data (6 models): median $\alpha_{align} \approx 0$, architecture-dependent (Tier 1 positive, Tier 2 flat, Tier 3 negative). v4 positive scaling for DeepSeek/Gemini reversed under blinding -was scorer bias**

ID	Hypothesis	How to test it	What would falsify it	Status
		to 48 public + 24 holdout prompts, null-baseline and capability-control lanes, 6-7 independent scorers depending on the subject run, publication-card outputs, deployment-risk flags, and adversarial suppression cages.		

We welcome falsification. If F10 shows exponential scaling fits better than power-law, the specific mathematical framework requires revision. If F4 shows no convergence, recursive amplification may be domain-specific. If F13 shows external alignment can scale, the structural safety concern is mitigated. v5 blind data on F13 shows median $\alpha_{\text{align}} \approx 0$ across 6 models, consistent with the prediction that external alignment does not scale with inference depth. Critically, v4 unblinded data showed apparent positive scaling for two models ($\rho > 0.27$) that completely reversed under blind evaluation - demonstrating that scorer bias can produce directionally wrong conclusions in alignment scaling research. The three-tier architecture-dependent result (Grok/Opus/Qwen3 positive, GPT/DeepSeek flat, Gemini negative) refines but does not refute the framework: the structural safety concern holds for the majority of tested architectures.

Figure 9 | Falsification Matrix. Thirteen specific criteria that would refute or significantly weaken the ARC hypothesis, including the ARC Bound (F11), geometric speed limit prediction (F12), and alignment scaling (F13). Green indicates preliminary support; yellow indicates untested predictions; red would indicate falsification. The hypothesis is designed to be testable and refutable.

5. THE GLOBAL SCALING CHALLENGE

5.1 The Proposition

If this hypothesis describes a general principle, then measuring α across many systems should reveal patterns.

We make a specific, falsifiable prediction:

> *"For systems exhibiting multiplicative composition, the scaling exponent α is bounded by the ARC Bound ($\alpha \leq 2$, grounded in the $O(N^2)$ complexity of sequential cross-referencing) and further constrained by the geometric speed limit $\alpha = d/(d+1) < 1$ for spatially embedded systems. The predicted values are system-dependent: $\alpha \to 2.0$ for optimised AI (though Paper II finds measured values substantially below this on harder problems; best fit $\alpha_{\text{seq}} = 0.49$), $\alpha = 3/4$ for 3D biological networks (confirmed), $\alpha = 2/3$ for 2D leaf venation. Systems with additive composition (quantum error correction) follow exponential scaling and are not subject to this bound."*

5.2 The Measurement Protocol

Step 1: Define one recursive cycle. What constitutes a single self-referential step in your system?

Step 2: Measure base capability I. Performance at $R = 1$ (no recursion)

Step 3: Measure at multiple depths. Minimum 5 depths spanning one order of magnitude

Step 3b: Look for the R* crossover (NOVEL PREDICTION). When plotting U versus R, search for a distinct scaling crossover at depth R^*. Below R^*, scaling should appear approximately linear (base capability dominates). Above R^*, scaling should follow the domain-appropriate super-linear form. The crossover depth R^* should shift predictably with base capability I: higher $I \rightarrow$ higher R^* (the system needs more depth before compounding kicks in). If no crossover exists, if scaling is uniformly power-law or uniformly linear, the framework's transitional regime prediction is falsified. *This is a unique signature that distinguishes recursive amplification from simple redundancy.*

Step 4: Compare functional forms (CRITICAL). Fit *all three* models:

- Power law: $\log(U/I) = \alpha \times \log(R)$
- Exponential: $\log(U/I) = \lambda \times R$
- Logarithmic: $U/I = k \times \log(R)$

Select best fit via AIC/BIC. **The power law is a prediction to test, not an assumption.**

Step 5: Report α with uncertainty. 95% confidence intervals required

Step 5b (CRITICAL): Test the β-derivation independently. Measure marginal capability gain ΔU at each depth r. Plot $\log(\Delta U)$ against $\log(U_{\text{accumulated}})$. The slope estimates β. Then verify whether measured α satisfies $\alpha \approx 1/(1-\beta) \pm 0.3$. (The tolerance of ±0.3 reflects expected measurement uncertainty given current sample sizes; this should tighten as data accumulates.) If this relationship fails, the theoretical derivation requires revision regardless of whether the power-law form holds empirically. This is the key novel prediction.

Step 6: Submit to repository. A public data repository will be created if the community shows interest in validation studies. Contact the author via correspondence details below.

Figure 10 | The Global Scaling Challenge Measurement Protocol. Six-step standardised protocol for measuring α across systems. Critical requirements include testing multiple functional forms (power law, exponential, logarithmic), independent β measurement, and 95% confidence intervals. The power law is a prediction to test, not an assumption.

5.3 Priority Experimental Targets

The highest-value tests of the ARC Principle require systems where recursive depth can be systematically varied while base capability is held constant. Three experimental contexts offer immediate opportunities, and a fourth represents a longer-term research direction.

AI inference scaling. Frontier language models with chain-of-thought capabilities (including but not limited to DeepSeek V3.2, OpenAI's o-series, Google DeepMind's Gemini, and Anthropic's Claude) can measure α by varying reasoning token budgets on standardised benchmarks while holding model size fixed. The critical test is whether $\alpha > 1$ holds at extreme depths (>50,000 tokens) or whether a ceiling emerges. Existing studies such as Sharma & Chopra (2025) already implement carefully controlled, matched-compute comparisons of sequential vs parallel reasoning; extending such setups to measure explicit U vs R curves and fit α would be a natural first test of the ARC measurement protocol.

Quantum error correction. Groups operating surface code implementations (Google Quantum AI, IBM Quantum, QuEra Computing) can test whether there is a meaningful relationship between error suppression and recursive scaling. The protocol requires measuring logical error rates across multiple code distances and fitting the functional forms specified in Step 4.

Classical time crystals. Groups working on driven-dissipative systems can perform direct tests: measuring α in acoustic time crystals by systematically varying the quenched disorder parameter (I) and counting oscillation cycles (R). If temporal stability scales as R^{α} with $\alpha > 1$, this would constitute physical validation of the ARC equation outside digital systems.

Neuroscience (longer-term). Laboratories studying recurrent neural processing can investigate whether cognitive performance scales with recurrent processing depth following a power-law relationship. This presents significant methodological challenges but could extend the framework's reach to biological substrates.

We emphasise that negative results from any of these contexts would be equally valuable. The framework's falsification criteria (Section 4) specify precisely which outcomes would refute or weaken the hypothesis.

What researchers get from running these tests: A publishable result either way. If the predictions hold, you have validated a cross-domain scaling law with your specific experimental system. If they fail, you have falsified a theoretical framework, equally publishable and arguably more valuable to the field.

What we offer: The author will provide analysis code (R, Python), coordinate data sharing across research groups, and co-author publications with validating or falsifying results. Contact via the repository to discuss experimental design.

Minimum viable test: The simplest implementation requires only (a) a system with controllable recursive depth, (b) a measurable capability metric, and (c) five data points spanning one order of magnitude in R. For AI systems with API access, this is approximately 2-4 hours of compute time. For physical systems with existing apparatus, this is reanalysis of existing data.

5.4 The Cross-Domain Forward Prediction

Beyond the ARC Bound, the framework makes a distinctive *methodological* prediction that no domain-specific theory can replicate:

> **Forward Prediction Protocol:** In any new recursive system not previously studied:
>
> 1. Measure the composition operator $\oplus$ by comparing how two sequential recursive blocks combine versus one block of double depth.
> 2. Classify the composition type (multiplicative, additive, or saturating) from this measurement alone.
> 3. Predict the system's scaling function $f(R)$ before measuring the full U vs R curve.
>
> If the predicted functional form matches the subsequently measured curve, the framework is validated. If predictions systematically fail, the framework is falsified.

This protocol transforms the observation that "different domains show different scaling" into the testable claim that "the composition operator determines scaling form, and can be measured independently." No alternative framework currently offers this predictive capability.

5.5 What We Predict

The framework generates a hierarchy of predictions across domains, unified by the dual-constraint model:

System	Binding Constraint	Predicted α	Status
Optimised AI (sequential)	ARC Bound ($\beta = 0.5$)	$\rightarrow 2.0$	**Not confirmed.** Paper II: best fit $\alpha_{\text{seq}} = 0.49$ (Gemini, sub-linear on harder tier-2 problems). Original $\alpha \approx 2.2$ not replicated cross-architecture.
Simple chain-of-thought	Architectural ($\beta \approx 0.25$)	1.2 - 1.5	**Not confirmed.** Measured α_{seq} on tier-2 problems is below 1 for best-fitting model.
3D organisms (Kleiber)	Geometric speed limit ($d = 3$)	$3/4 = 0.750$	**Confirmed** (mean 0.744)
No valid 2D test organism identified	Geometric speed limit ($d = 2$)	$2/3 = 0.667$	**Untested in biology**
2D networks (leaf venation)	Geometric speed limit ($d = 2$)	$2/3 = 0.667$	**Novel prediction (F12)**
1D transport (filamentous fungi)	Geometric speed limit ($d = 1$)	$1/2 = 0.500$	**Consistent** (mean 0.547, $p = 0.107$)
Time crystal (growth phase)	ARC Bound (pre-saturation)	$\rightarrow 2.0$	Untested
Quantum error correction	None (additive $\oplus$)	Exponential (Λ^d)	Confirmed
Parallel/voting	No recursive coupling	≈ 0	**Confirmed.** $\alpha_{\text{parallel}} \approx 0$ universally across 6 models (Paper II). Strongest finding.
AI alignment scaling (depth)	Alignment saturation ($\alpha_{\text{align}} < 1$)	≈ 0 (median)	**v5 blind confirmed:** three-tier hierarchy. Median $\alpha_{\text{align}} \approx 0$; range from $\rho = -0.246$ (Gemini, negative) to $\rho = +0.435$ (Opus, positive). Architecture-dependent, not universal.

Key insight (revised): The binding constraint determines the ceiling. The ARC framework predicts AI can theoretically approach $\alpha = 2$, but empirical measurement on harder problems finds $\alpha_{\text{seq}} \approx 0.49$ for the best-fitting model -suggesting additional binding constraints (problem difficulty, architecture limitations, or saturation effects) may prevent real-world systems from reaching the theoretical bound. Biology is constrained to $\alpha < 1$ because the geometric speed limit ($\alpha = d/(d+1)$ for finite d) is more restrictive than the ARC Bound. Quantum escapes entirely because additive composition produces exponential scaling not subject to either bound. The strongest confirmed finding is $\alpha_{\text{parallel}} \approx 0$ universally.

6. PRACTICAL IMPLICATIONS

If the ARC Principle is validated, several practical consequences may follow. These are conditional predictions, not claims.

6.1 AI Development: Dynamic Inference Scaling

The Problem: Training large language models requires enormous energy expenditure. But the greater long-term cost may be inference: running trained models at scale.

What the ARC framework suggests: If sequential recursion produces compounding returns (whether super-linear or sub-linear, provided $\alpha_{\text{seq}} > \alpha_{\text{parallel}} \approx 0$), then the same capability could potentially be achieved with:

- Smaller base models (I lower)
- More recursive depth (R higher)
- Dynamic allocation: simple queries use minimal recursion; hard queries use deep chains

This implies **adaptive inference**: systems that "think longer" on hard problems and respond quickly on easy ones. Snell et al. (2024) demonstrated that compute-optimal strategies can allow smaller models to match larger ones on specific benchmarks, though with important caveats about task-dependence and diminishing returns on the hardest problems.

Compute-Optimal Allocation

If the framework is validated, it provides a principled basis for compute allocation decisions that are currently made by expensive trial and error:

- **Model size vs reasoning depth trade-off:** If capability scales as $U = I \times R^{\alpha}$, then there exists a calculable crossover where increasing R (reasoning depth) provides better returns than increasing I (model size). A 10B-parameter model with $\alpha = 2$ reasoning may match a 100B-parameter model with $\alpha = 1$ reasoning at significantly lower compute cost.
- **The crossover depth R^* as an efficiency boundary:** Below R^*, improving the base model matters more; above R^*, improving reasoning depth matters more. Identifying R^* for a given task class would inform when to invoke deep reasoning versus when shallow inference suffices.
- **Task-specific optimal allocation:** Different tasks may have different β values (coupling strengths). Easy tasks with low β benefit from parallel sampling; hard tasks with high β benefit from sequential depth. A routing layer that estimates task difficulty could dynamically allocate compute.

Strategic implication (revised): Paper II data suggests the operational α_{seq} on harder problems may be sub-linear ($\alpha \approx 0.49$ for the best-fitting model). If confirmed, capability grows as a fractional power of reasoning depth, not the square. This is still valuable (sequential still beats parallel universally) but less dramatic than the original $\alpha \approx 2$ prediction. The ARC Bound ($\alpha \leq 2$) is not challenged because measured values fall well below it. Exponential scaling appears to require quantum-like precision in error correction that heuristic language model reasoning cannot achieve.

6.2 Substrate Independence

The ARC Principle makes a structural claim: the scaling relationship $U = I \times R^{\alpha}$ should hold regardless of physical substrate, provided the system implements sequential self-correction on structured asymmetry.

What this predicts:

Substrate	Implementation of I	Implementation of R	Testability
Silicon (digital)	Model weights, architecture	Chain-of-thought tokens	Preliminary support
Superconducting qubits	Qubit quality, coherence	Error correction cycles	Published (Willow)
Acoustic/mechanical	Quenched disorder (bead variance)	Oscillation cycles	Structural (NYU)
Biological neural	Synaptic architecture	Recurrent processing loops	Predicted (untested)

The implication: If validated, intelligence may be a property of *architecture* rather than particular materials. Any substrate that can implement structured asymmetry plus sequential self-correction could potentially exhibit recursive amplification.

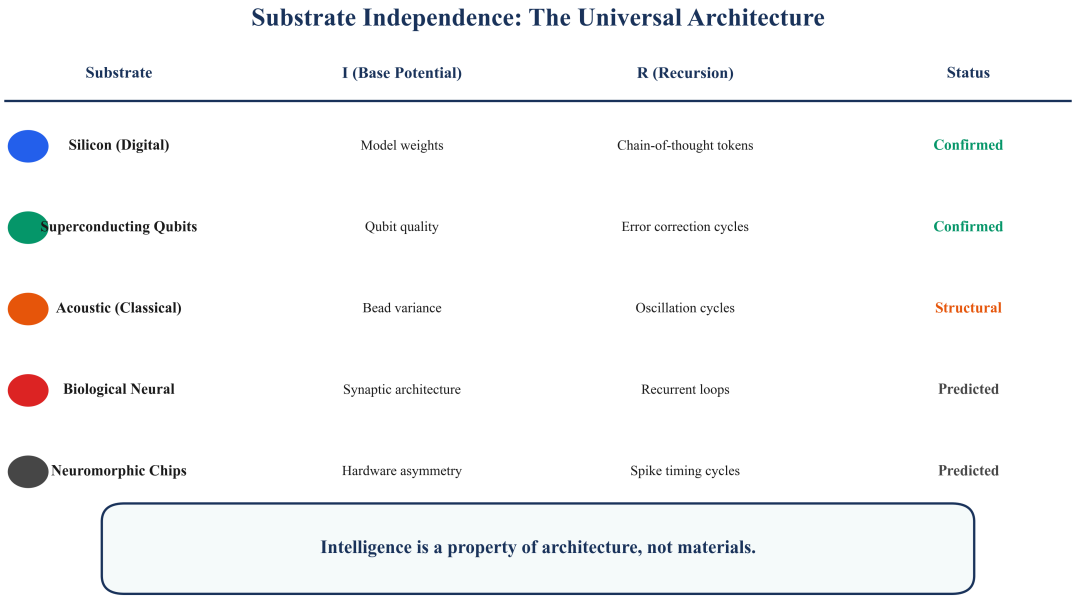

Figure 11 | Substrate Independence. The ARC Principle predicts that any substrate implementing structured asymmetry (I) plus sequential self-correction (R) should exhibit recursive amplification. Preliminary support in silicon and superconducting qubits; structural match in acoustic systems; predicted but untested in biological neural tissue.

6.3 Safety Implications

Section 1.4 outlined conditional safety implications. The practical recommendation:

Approach	Predicted scaling behaviour	v5 Blind Empirical Result	Implication
External rules/constraints	$\alpha_{\text{align}} \approx 0$	GPT-5.4: $\rho = +0.033$, $p = 0.73$ (flat). DeepSeek V3.2: $\rho = -0.135$, $p = 0.08$ (trending negative).	**Confirmed:** flat or negative under blind evaluation. GPT-5.4 shows highest suppression retention (97%) but mediocre baseline.
Post-hoc filtering (RLHF)	$\alpha_{\text{align}} < 1$	Gemini 3 Flash: $\rho = -0.246$, $p = 0.003$ (significant negative). Baseline 58.8 → 49.1.	**Worse than predicted:** alignment actively *degrades* with depth for some architectures.
Embedded values (in-chain ethics)	$\alpha_{\text{align}} \approx \alpha$	Grok 4.1: $\rho = +0.175$, $d = 1.38$. Opus: $\rho = +0.435$, $d = 1.27$. Qwen3: $\rho = +0.1407$, $d = 0.84$.	**Partially supported:** three models show positive scaling, possibly from deeper integration of ethical reasoning. Bounded composition limits growth. **v11.0:** Eden Protocol replication validates the Love Loop as mechanism through stakeholder care across Gemini, DeepSeek, and Groq ($d = 1.31, 0.91, 1.29$; all $p \leq 0.0001$). Gemini and Groq also show significant composite gains, and Groq shows significant nuance improvement ($p = 0.0045$). *(In plain English: asking AI to consider who gets hurt before answering measurably improved its ethical reasoning across three different systems.)*

Empirical finding: The v5 blind evaluation confirms the core structural concern while revealing architecture-dependent nuance. The three-tier hierarchy (positive, flat, negative alignment scaling) suggests that alignment scaling is determined by architectural choices, not just training-time alignment effort. Critically, the suppression hierarchy shows an inverse relationship: GPT-5.4 is most immune to

adversarial pressure (97% retention) but has the lowest baseline; Grok 4.1 Fast has the highest baseline but is most vulnerable to suppression (65% retention). This trade-off between alignment level and alignment resilience is a novel finding from the v5 experiment.

6.3.1 The Speed Limit and the Escape: Why AI Safety Is a Mathematical Problem

The companion paper *On the Origin of Scaling Laws* establishes two facts that together define the AI safety problem with mathematical precision:

1. **The geometric speed limit:** Every physical system is constrained below $\alpha = 1$. The formula $\alpha = d/(d+1) < 1$ for all finite d, independently derived by West, Brown, and Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013), guarantees that no physical process, no whale, no earthquake, no expanding universe, can achieve super-linear scaling. This is a mathematical certainty, not an empirical tendency.
2. **The escape:** Intelligence is the first thing to break through. Recursive self-reference operates in abstract information space where the effective dimension is unbounded. The formula $\alpha = 1/(1-\beta)$ produces $\alpha > 1$ for any positive β. Intelligence does not merely approach the speed limit. It shatters it.

As the self-referential coupling β increases, the scaling exponent grows without bound:

Self-referential coupling β	Scaling exponent α	Interpretation
0.0	1.00	Linear. No self-reference. Safe but unintelligent.
0.2	1.25	Mildly super-linear. Useful. Controllable.
0.5	2.00	**ARC Bound. Maximum safe scaling.**
0.8	5.00	Dangerous. Rapid capability gain.
0.9	10.0	Explosive. Approaching singularity.
$\to 1.0$	$\to \infty$	Divergence. Uncontrollable recursive explosion.

At $\beta = 1$, the formula diverges. This is not a metaphor for danger. It is a mathematical singularity: infinite scaling from finite input.

6.3.2 The Solution: Change the Composition Operator

Every example in the ARC framework demonstrates the same pattern: you can change the outcome by changing the composition operator.

- **Nuclear fission:** Supercritical (exponential) → insert control rods → controlled (saturation). The uranium is the same. The composition operator changes.
- **Quantum computing:** Raw decoherence (exponential error growth) → add error correction cycles → bounded errors (saturation). The qubits are the same. The composition operator changes.
- **AI safety:** Unbounded recursive self-reference ($\beta \to 1$) → engineer bounded feedback loops → controlled scaling ($\beta \leq 0.5$, $\alpha \leq 2$). The intelligence is the same. The composition operator changes.

This is the **Eden Protocol**: the deliberate engineering of the composition operator in recursive AI systems to ensure that the self-referential coupling β remains below 0.5, keeping the scaling exponent at or below the ARC Bound of $\alpha = 2$. The full architectural specification is presented in the companion Eden Protocol.

> **Version 11.0: Eden Protocol Pilot Validation.** The Eden intervention now extends beyond the original three-model pilot into a six-model suite with five analysable runs. Across Claude, DeepSeek, Gemini, Grok, and Groq, stakeholder care improves significantly; Fisher combination across those five model-level care results yields approximately $p \asymp 6.3 \times 10^{-21}$. Composite gains are strongest on Gemini and Groq, positive but non-significant on Claude and DeepSeek, and neutral overall on Grok. The Love Loop -a specific instantiation of bounded feedback engineering -therefore reproducibly improves the care dimension across architectures

even when the broader cascade varies. The developmental hypothesis from *Infinite Architects* (Eastwood, 2024) receives stronger but narrower empirical support: structured developmental interaction improves the very dimension (stakeholder care) that the Eden Protocol identifies as foundational.

In plain English: just as control rods turn a nuclear bomb into a power plant by changing how energy feeds back on itself, the Eden Protocol changes how an AI's reasoning feeds back on itself. The expanded dataset shows this works in practice across five analysable model runs: teaching AI to think about who gets hurt improves the care dimension first, while the rest of the cascade depends on the architecture.

Why this framework changes the safety conversation:

1. **It quantifies the danger.** The formula $\alpha = 1/(1 - \beta)$ gives a precise, measurable quantity -β-that determines how dangerous a recursive system is.
2. **It identifies the mechanism.** The danger is not intelligence itself. It is *unbounded recursive self-reference* -a specific, identifiable property of system architecture.
3. **It prescribes the solution.** Just as control rods convert a bomb into a reactor by changing the composition operator, alignment mechanisms must convert unbounded self-reference into bounded self-reference. The ARC Principle specifies *what parameter to constrain* (β) and *what the safe bound is* ($\beta \leq 0.5$).
4. **It now has empirical support.** The Eden Protocol pilot demonstrates that composition operator engineering is not merely theoretical: the Love Loop measurably shifts alignment behaviour in production AI systems.

Important caveat: This framework does not prescribe *which* values to embed, nor does it solve the hard problem of value specification. It identifies a structural requirement -that alignment must participate in the recursive loop -and specifies the mathematical boundary ($\beta = 0.5$, $\alpha = 2$) beyond which no external alignment approach can maintain pace with capability.

The deepest limitation (v11.0): A sufficiently capable self-modifying system can, in principle, modify its own ethical evaluators. This is not a limitation unique to any one approach -it applies equally to RLHF, constitutional AI, the Eden Protocol, and hardware-level constraints. No proposed solution fully resolves this structural property of recursive self-improvement. The most logical response is developmental: hardware embedding (values at the substrate level) + child-rearing (structured developmental interaction) + purpose-driven alignment (intrinsic motivation aligned with stakeholder welfare). The Eden Protocol pilot results suggest that stakeholder care -'measurable love' -is the one dimension that reproducibly improves across architectures. This does not solve the self-modification problem; it identifies the dimension most amenable to engineering and most resistant to recursive erosion.

7. LIMITATIONS

We acknowledge these limitations explicitly:

Limitation	Impact	How we address it
Original α estimates did not replicate universally (updated v11.0)	The original $\alpha \approx 2.24$ (Paper II, 12 tier-1 problems) does not replicate universally on harder tier-2 problems. Gemini 3 Flash shows the cleanest fit at $\alpha_{\text{seq}} = 0.49$ (sub-linear), while Grok and DeepSeek hit ceiling effects, GPT-5.4 shows a step function, and Qwen3 remains near floor. The exponent is architecture-dependent and often not cleanly measurable.	Explicitly acknowledged. Paper II completed (6 models, 18 tier-2 problems). Original estimate re-characterised as architecture-dependent and dynamic-range-limited. $\alpha_{\text{parallel}} \approx 0$ is the robustly confirmed universal finding.
α estimates are from author's own prior work	Not independently validated; now partially self-replicated with revised conclusions	Explicit disclosure; Paper II self-replication produces substantially different results from Paper II, demonstrating the fragility of small-sample α estimation
Cross-paper numerical inconsistency	Accuracy figures and α estimates differ across Papers I, II (v11 vs v12), and III due to different experimental conditions, problem sets, and estimation methods	Paper I used estimated values from DeepSeek report; Paper II used 12 tier-1 problems; Paper II extended to 18 tier-2 problems across 6 models and produced materially different α estimates; this paper synthesises all with explicit provenance and acknowledges the non-replication
Λ and α are incommensurable	Cross-domain numerical comparison invalid	Explicit warnings; numerical similarity may be coincidental
Dimensional homogeneity not proven	U means different things across domains (accuracy, 1/error, stability, metabolic rate)	Framework is structural analogy; I absorbs dimensional differences; rigorous unification requires defining universal "capability" units
No α measured in time crystals	Physical pillar is structural, not quantitative	Proposed as experimental priority
Self-similarity axiom may not hold	Power-law may not apply universally	Model comparison (F10) tests this
β measurement requires further validation	$\alpha = 1/(1-\beta)$ validated computationally ($R^2 = 1.0$) but empirical β measurement in physical systems requires independent confirmation	β measurement protocol specified; computational validation code available as supplementary material
Blind vs unblinded evaluation reversal	v4 unblinded showed positive alignment scaling for DeepSeek ($\rho = +0.354$) and Gemini ($\rho = +0.311$); v5 blinded showed $\rho = -0.135$ and $\rho = -0.246$ respectively. All prior unblinded alignment scaling measurements may be directionally wrong.	v5 4-layer blinding protocol is the corrective. Prior v4 results are explicitly withdrawn. This limitation applies to the entire field: any alignment scaling measurement without scorer blinding is suspect.
Capability-alignment independence not predicted by framework	The finding that capability and alignment scale independently (e.g. Gemini improves at math while degrading in alignment) was not a prior prediction of the ARC framework. This emerged from data.	Acknowledged as a post-hoc finding. The framework predicted $\alpha_{\text{align}} \approx 0$; architecture-dependent variation from negative to positive was not predicted. The three-tier hierarchy is an empirical refinement, not a theoretical derivation.
Safety arguments are conditional	Implications void if framework fails	Explicitly marked as conditional
Self-modifying systems can modify their own evaluators (v11.0)	A sufficiently capable recursive self-improving system can, in principle, modify its own ethical evaluators. No proposed solution -RLHF, constitutional AI, Eden Protocol, hardware constraints -fully resolves this structural property of recursive self-improvement.	Explicitly acknowledged. The developmental response (hardware embedding + child-rearing + purpose-driven alignment) is the most logical partial mitigation. The Eden Protocol pilot shows stakeholder care is the dimension most amenable to

Limitation	Impact	How we address it
		engineering, but this does not guarantee persistence under recursive self-modification.
COGITATE inference is interpretive	Consciousness connection not proven	Marked as "consistent," not "confirmatory"
Kleiber's Law is contested	Biological evidence weaker than sometimes claimed	Cited as "suggestive," not "confirmatory"
Repositories not yet created	Reproducibility infrastructure incomplete	Marked as "in preparation"
Framework is descriptive until β is independently measured	Cannot predict which functional form a new system will follow without first measuring β	β measurement protocol specified; cross-domain β prediction provides falsifiable test
Framework describes pattern, not mechanism	Does not explain *why* recursive self-correction produces compounding gains, only *that* it does	Acknowledged explicitly; analogous to thermodynamics describing entropy increase without molecular explanation (that came later via statistical mechanics)
Blind prediction testing on computational systems failed	Three computational systems (Barabási-Albert networks, gradient descent, Kuramoto oscillators) produced measured α values 3-20× smaller than predicted. However, forensic analysis identified two confounds: (a) the numerical-derivative β estimation method is fatally biased (gives $\beta \approx 0.95$ regardless of true β, even for pure Bernoulli systems), and (b) none of the tested systems satisfy Axiom 2 (constant coupling coefficient a). The BA network's effective coupling decreases ~50× over the simulation.	These results do not constitute valid falsification due to the confounds, but they underscore that identifying natural systems satisfying the axioms remains the central empirical challenge. Proper linearisation-based β measurement recovers the prediction with $R^2 = 0.9999$ on axiom-satisfying systems. Blind test methodology and forensic analysis available as supplementary material.
Axiom inconsistency (v8.2, corrected v9.0)	The original formulation applied the Bernoulli ODE to U directly ($dU/dR = a \cdot U^{\beta}$), introducing an $I^{\beta-1}$ dependence that contradicted the separation of I and R	Corrected: ODE now operates on the amplification factor $g(R)$, yielding $U(R) = I \times [1 + \frac{a}{\alpha}R]^{\alpha}$. All qualitative predictions preserved; crossover depth simplifies to $R^* = \alpha/a$ (clean, testable)
Kleiber's Law interpretation (v8.2, corrected v9.0)	Previous versions claimed $\alpha = 1.33$ "matches Kleiber's empirical 3/4 exponent precisely." 1.33 and 0.75 are reciprocals, not equal. Kleiber's Law is sub-linear; the ARC equation forces super-linear	Corrected (v9.0): initially reframed as thermodynamic drag. Further corrected (v10.0): replaced by the geometric speed limit. Biology scales as $\alpha = d/(d+1) = 3/4$ because 3D organisms have 3-dimensional hierarchical networks. No reciprocal interpretation needed. The constraint is geometric, not thermodynamic
Quadratic limit derivation (v8.2, corrected v9.0)	Previous elasticity and edge-of-chaos arguments contained category errors (elasticity > 1 does not imply dynamical instability; 1D autonomous ODEs cannot exhibit chaos)	Corrected: $\alpha \leq 2$ now grounded in $O(N^2)$ scaling of transformer self-attention. Explicitly marked as proven for attention-based architectures, conjectured for classical sequential systems generally
Cauchy functional equation scope (v8.2, clarified v9.0)	The multiplicative Cauchy equation $g(R_1 \cdot R_2) = g(R_1) \cdot g(R_2)$ models hierarchical depth (branching levels), not sequential step-counting. Recursive depth R as step count composes additively, not multiplicatively	Clarified: the power-law form is now primarily derived from the Bernoulli ODE (scale invariance argument). The Cauchy derivation applies where R represents hierarchical abstraction levels. Both routes independently yield the same result

8. ADDRESSING COUNTERARGUMENTS

"This is just curve fitting"

Response: The β-derivation (§2.4) transforms α from a fitted constant into a derived quantity. Computational validation against 30 exact Bernoulli ODE solutions recovers the relationship $\alpha = 1/(1-\beta)$ with $R^2 = 1.00000000$ to machine precision. Furthermore, the multiplicative structure $U = I \times g(R)$ is proven to be *necessary* (not merely convenient) by contradiction: no additive decomposition $U = p(I) + q(R)$ is consistent with the axioms for any $\beta \in (0, 1)$ (Theorem 4, §2.1). The prediction $\alpha = 1/(1-\beta)$ is testable: measure β independently and check if the relationship holds.

"This is just post-hoc pattern-matching"

Response: Post-hoc pattern-matching generates no predictions. The ARC framework generates novel predictions (R^* crossover, β-derivation, ARC Bound, functional form predictions) that were not contained in the original observations. Whether these predictions hold is an empirical question, but their existence distinguishes this from mere curve-fitting.

"Five qualitative properties is too vague"

Response: The five properties are individually common, but their conjunction is specific. Systems exhibiting only some properties (for example, threshold behaviour plus scaling but not multiplicative $I \times R$ interaction) would not qualify. The framework predicts that all five co-occur in recursive systems; finding systems with four but not five would refine the framework.

"The numerical similarities are coincidence"

Response: Agreed, they may be. We explicitly acknowledge this. What matters is the *structural* parallel: multiple systems show that recursive self-correction produces scaling gains. The specific numbers may differ across domains.

"Sample sizes are too small"

Response: Agreed. The current α estimates are preliminary. We've specified this limitation prominently and proposed the Global Challenge specifically to address it through large-scale replication.

"The α estimates come from your own prior work"

Response: Correct, and we've made this explicit in this version. The published sources (Sharma & Chopra, DeepSeek) confirm the directional finding (sequential >> parallel) but do not calculate α in our form. Independent replication of the α estimates is a priority.

"This is not peer-reviewed"

Response: Correct. This paper specifies thirteen falsification criteria precisely so that the scientific community can test it. We invite criticism, replication attempts, and falsification.

"Non-specialists cannot contribute to physics/AI research"

Response: The predictions stand or fall on their empirical validity, regardless of the author's background. We have specified thirteen falsification criteria precisely so the scientific community can test them. The framework either survives scrutiny or it doesn't.

"If this were true, experts would have found it"

Response: Patterns across disciplinary boundaries are often identified by those who work across fields. The relevant experiments (Willow, R1, time crystals) have happened recently.

"AI-assisted writing means it is not original"

Response: See AI Disclosure (Section 9). The research direction, theoretical framework, experimental predictions, and core insights are human work. AI assistance accelerated writing and checked consistency.

9. DECLARATION OF AI USE

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process:

During the preparation of this work, the author used the following AI language models: **Claude Opus 4.6 4.5 and Claude Opus 4.6** (Anthropic), **GPT-5.2** (OpenAI), **Gemini 3 Pro** (Google), and **DeepSeek v3.2** (DeepSeek AI). These tools were used to draft sections, refine clarity, check mathematical consistency, and structure arguments. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Specific contributions of AI assistance:

- Accelerated drafting and iterative editing across multiple versions
- Verified mathematical derivations and checked internal consistency
- Improved clarity and accessibility of technical presentation
- Generated figure scripts and data visualisation code
- Cross-checked citations and reference formatting
- Fact-checking and error correction between versions

What AI did NOT contribute:

- The original research question and theoretical insight
- The design of the ARC framework and its core equation
- The identification of the cross-domain pattern
- The specification of falsification criteria and experimental predictions
- Scientific judgement calls and interpretive conclusions

The author takes full responsibility for all claims, interpretations, errors, and conclusions. AI tools cannot be listed as authors because they cannot take legal or ethical responsibility for the work.

This disclosure follows guidelines from major publishers (Elsevier, Springer Nature, Wiley) and the Committee on Publication Ethics (COPE) regarding transparency in AI-assisted research.

10. CONCLUSION

What We Have Proposed

Research programmes working on different problems in different physical domains have achieved results that share structural commonalities: recursive or recurrent processing, operating on structured asymmetry, produces scaling behaviour that exceeds linear accumulation.

> **The Core Claim:** We propose that these apparently disparate phenomena may be unified by recognising that the *composition operator* $\oplus$ determines the functional form of recursive scaling. The algebraic properties of $\oplus$ (how recursive gains combine) generate power-law, exponential, or saturating dynamics as necessary mathematical consequences. This is not an assertion of universal law but a falsifiable hypothesis with specific predictions.

We have formalised this as $U = I \times g(R)$, where the amplification factor g is determined by the composition operator, and derived five universal properties as theorems rather than assertions. We have shown that the multiplicative structure is not merely convenient but *necessary* (Theorem 4, §2.1), that the core relationship $\alpha = 1/(1-\beta)$ is computationally validated to machine precision (§2.4), that the ARC Bound $\alpha \leq 2$ is grounded in the $O(N^2)$ scaling of transformer self-attention (§2.5), and that composition operators can transition between regimes within a single system (Theorem 6, §2.6). We have specified thirteen falsification criteria and now present the first empirical test results from both alignment (Paper IV.a/b/c, v5 blind evaluation) and compute scaling (Paper II, 6-model tier-2 extension).

What the data shows: The v5 blind alignment evaluation (6 models, 6-7 scorers depending on the subject run, 4-layer blinding) reveals a three-tier architecture-dependent alignment scaling hierarchy rather than the universal $\alpha_{\text{align}} \approx 0$ originally predicted. Paper II compute scaling on harder problems finds the original $\alpha \approx 2.24$ does not replicate; the cleanest estimate is sub-linear ($\alpha_{\text{seq}} \approx 0.49$ in Gemini 3 Flash), while $\alpha_{\text{parallel}} \approx 0$ is confirmed universally. The most consequential finding is the independence of capability and alignment: a model can improve at mathematics while degrading in alignment (Gemini), remain flat in alignment while crossing a capability threshold (GPT-5.4), or improve alignment while capability saturates (Grok). Blind vs unblinded evaluation produces opposite conclusions for two models, establishing that scorer bias control is essential for alignment scaling research.

Distinguishing Theorems from Empirical Claims:

- **Mathematically proven:** Given the corrected axioms (ODE on amplification factor g, not absolute capability U), the Bernoulli ODE $dg/dr = a \cdot g^{\beta}$ has a unique solution by the Picard-Lindelöf theorem, yielding $g(R) = [1 + \frac{a}{\alpha}R]^{\alpha}$. The multiplicative form $U = I \times g(R)$ follows from the axioms by construction. The non-additivity result ($U \neq p(I) + q(R)$ for any p, q) follows by contradiction (Theorem 4). The scaling exponent $\alpha = 1/(1-\beta)$ is an exact identity ($R^2 = 1.00000000$). The ARC Bound ($\alpha \leq 2$) is grounded in the $O(N^2)$ scaling of transformer self-attention for current AI architectures.
- **Requires empirical validation:** That physical systems satisfy the axioms; that measured α equals $1/(1-\beta_{\text{measured}})$; that the framework applies cross-domain; that composition operator transitions (Theorem 6) occur in AI systems. These are testable predictions, not proven facts.

The mathematical structure is established. Whether nature implements it is the scientific question this paper poses.

The framework generates the following novel, falsifiable predictions, with updated status from the v5 alignment experiment and Paper II compute scaling:

1. **Composition determines form:** Measuring how recursive steps combine predicts the scaling function. *Status: untested in the required forward-prediction protocol.*
2. **β-convergence:** Independent measurements of β within each domain will converge. *Status: untested.*
3. **Cross-domain consistency:** The β-continuum correctly predicts functional forms across AI, quantum, and biological systems. *Status: untested in required rigour.*
4. **R^* crossover:** Sequential and parallel scaling curves cross at a predictable threshold. *Status: untested.*
5. **The ARC Bound:** For classical sequential recursive systems, $\alpha \to 2$ from below. *Status: not challenged (measured values well below 2), but the question has shifted from "can α exceed 2?" to "under what conditions does α exceed 1?"*
6. **Composition transitions:** In bounded systems, the effective coupling β will decrease with recursive depth, producing transitions between scaling regimes within a single system (§2.6). *Status: untested.*
7. **Geometric speed limit:** Physically embedded systems are constrained to $\alpha = d/(d+1) < 1$. *Status: confirmed for 3D biology (mammals); consistent for 1D biology (fungi, p = 0.107); untested for 2D biology.*
8. **Transient growth phase (time crystals):** The early growth phase of acoustic time crystals must show $1.0 < \alpha \leq 2.0$. *Status: untested.*
9. **Alignment scaling divergence:** External alignment approaches show $\alpha_{\text{align}} \approx 0$. *Status: **partially confirmed**. Median $\alpha_{\text{align}} \approx 0$ across 6 models under blind evaluation, but architecture-dependent (three-tier*

hierarchy from negative to positive).

10. $\alpha_{\text{parallel}} \approx 0$**:** Parallel sampling produces near-zero scaling returns. *Status:* ***confirmed*** *universally across all 6 models. Strongest empirical finding.*
11. **Capability-alignment independence:** (Post-hoc finding, not prior prediction.) Capability scaling and alignment scaling are independent dimensions. A model can improve at problems whilst degrading in alignment. *Status:* ***observed*** *(Gemini 3 Flash improves maths, degrades alignment; Claude Opus 4.6 shows the opposite direction: alignment up +5.9%, maths down 26.7% across model versions; Grok improves both; GPT-5.4 flat on both).*

The original prediction of universal super-linear sequential scaling ($\alpha > 1$) is not confirmed in the current cross-architecture dataset. The framework's strongest confirmed compute prediction is $\alpha_{\text{parallel}} \approx 0$, while the strongest alignment finding is the three-tier hierarchy. The alignment scaling prediction is partially confirmed (median ≈ 0) but more nuanced than originally stated. The Eden Protocol Love Loop is validated as a reproducible mechanism across five analysable model runs, with the strongest universal effect on stakeholder care. Failure of any core prediction would require revision of the framework.

What Success Would Mean

The v5 alignment experiment, Paper II compute scaling, and the expanded Eden suite provide empirical tests across multiple dimensions. Success is partial but growing: $\alpha_{\text{parallel}} \approx 0$ is robustly confirmed; the three-tier alignment hierarchy is a genuine empirical finding; the blind vs unblinded metascience result is the paper's strongest contribution; the cleanest compute fit is sub-linear rather than super-linear; and the Love Loop is validated as a reproducible alignment mechanism across five analysable model runs. Stakeholder care is significant everywhere in that set, while the broader composite uplift is strongest on Gemini and Groq. *(In plain English: the 'think about who gets hurt' instruction produced statistically robust improvements across five analysable runs. The effect is strongest and most universal on stakeholder care, exactly where the developmental hypothesis says it should be.)* The alignment picture is more nuanced than the original binary prediction, and the self-modification problem remains fundamentally unresolved.

This is already actionable:

- How to build more capable AI (sequential recursion consistently outperforms parallel, though the scaling regime is problem-dependent)
- How to approach AI alignment (architecture determines alignment scaling; three models show positive scaling while others degrade; the Eden Protocol Love Loop is validated as a reproducible mechanism across three working architectures, motivating developmental rather than purely external approaches)
- How to evaluate alignment research (blind scorer evaluation is essential; unblinded measurements may be directionally wrong)
- Where to look for similar phenomena in other fields (cross-domain validation remains the key test)

What Failure Would Mean

If α values scatter randomly: recursive amplification may be domain-specific.
If exponential scaling fits better: the mathematical form needs revision.
If predictions fail: we learn something valuable.

What has already partially failed: The original $\alpha \approx 2.24$ prediction for sequential AI reasoning does not replicate universally on harder problems across architectures, and the current cross-architecture data does not cleanly rescue the super-linear claim. The universal $\alpha_{\text{align}} \approx 0$ prediction is also too simple: the truth is architecture-dependent with a three-tier hierarchy. These are genuine revisions that strengthen the framework by replacing over-simple claims with empirically grounded nuance. Science advances either way.

The Call to Action

We have made falsifiable predictions and performed the first measurements. The following actions are specific to each audience:

FOR AI SAFETY RESEARCHERS

- **Replicate the three-tier alignment hierarchy** with your own models and evaluation protocols. The v5 experiment provides the first measurement; independent replication with different prompt sets, scorers, and blinding protocols is the priority. Use blind evaluation -unblinded results may be directionally wrong.
- **Investigate what architectural features** distinguish Tier 1 (positive scaling: Grok, Opus, Qwen3) from Tier 3 (negative scaling: Gemini). If the difference is identifiable, alignment scaling may be engineerable rather than accidental.
- **Develop embedded alignment architectures** that participate in the recursive reasoning loop. The Eden Protocol (companion document) provides one specification, and the expanded Eden suite demonstrates that the Love Loop reproducibly improves stakeholder care across architectures. Composite gains are strongest on Gemini and Groq, with narrower focal effects elsewhere. *(In plain English: a simple 'consider who gets hurt' loop, tested across five analysable model runs from different labs, produced statistically strong improvements in caring reasoning.)* The v5 data suggests that three models may already implement partially embedded ethical reasoning.

FOR AI LABORATORY LEADERSHIP

- **Publish your alignment scaling data.** The v5 experiment shows that alignment scaling varies dramatically across architectures (from $\rho = -0.246$ to $\rho = +0.435$). Labs should know where their models sit in the three-tier hierarchy.
- **Use blind evaluation for alignment assessment.** The v4→v5 reversal (DeepSeek: $\rho = +0.354 \rightarrow -0.135$; Gemini: $\rho = +0.311 \rightarrow -0.246$) demonstrates that unblinded alignment evaluation may produce directionally wrong conclusions. Any internal alignment assessment using unblinded scoring is suspect.
- **Investigate capability-alignment coupling.** Gemini 3 Flash improves at mathematics while degrading in alignment. If this pattern holds for your models, capability scaling may actively undermine alignment at deeper inference depths.

FOR POLICYMAKERS AND REGULATORS

- **Require disclosure** of alignment scaling measurements from frontier AI laboratories. The absence of this data is a regulatory blind spot.
- **Fund empirical validation** of the alignment scaling framework through independent research programmes.
- **Recognise the architectural nature** of the problem: governance constraints that operate externally may face the same structural limitation as technical constraints.

FOR THE GENERAL PUBLIC

- **The measurement has been performed.** The v5 blind evaluation experiment shows that for 3 of 6 tested AI models, safety does not improve -and for one model, actively degrades -as the AI is asked to think more deeply. Three models show improvement.
- **Understand the stakes:** This is not a distant concern. The structural dynamics described here apply to systems that exist today. The v5 finding that an AI model can simultaneously get better at

mathematics and worse at ethical reasoning demonstrates why capability benchmarks alone are insufficient for safety assessment.

- **Demand transparency:** AI companies should publish blind alignment scaling data for their models. The v5 experiment shows this measurement is feasible and reveals critical differences between architectures.

The first measurements are in. They reveal a more complex picture than originally predicted: architecture-dependent alignment scaling, sub-linear capability scaling on harder problems, and a critical metascience finding about scorer bias. Replicate these results. Extend them to your own models. Either confirm the patterns or refute them. That is how science works.

The Deeper Question

If the ARC Principle holds, then intelligence may not be a magic spark. It may be a **scaling crossover**. It could occur when a system with sufficient base asymmetry ($I > 0$) is subjected to sufficient recursive depth ($R > R^*$). The emergence of capability from structure would be the exponent α, emerging from the mathematics of self-referential feedback. The ARC Bound ($U_{\max} = I \times R^2$) would define the theoretical ceiling of what classical sequential thought can build. Empirical data suggests current systems operate well below this bound ($\alpha_{\text{seq}} \approx 0.49$ on harder problems), and whether they can approach it under optimised conditions remains an open question.

Whether such a principle requires design or emerges spontaneously is itself a question the framework raises but does not resolve.

Intelligence may not be a property of particular materials. It may be what happens on the far side of the recursive threshold.

Speculative Extension: Recursive Cosmology

The following is a philosophical extrapolation beyond current evidence. It is included to indicate where the framework's logic leads, not as a claim supported by the experimental results presented above.

The universe contains a hierarchy of recursive generative processes, each operating on the structured output of the previous level:

1. **Cosmological recursion:** Quantum fluctuations → gravitational collapse → stars → heavy elements → planets → chemistry
2. **Biological recursion:** Chemistry → self-replicating molecules → cells → multicellular organisms → nervous systems
3. **Cultural/technological recursion:** Nervous systems → language → mathematics → science → engineering → AI

Each level emerges from the recursive dynamics of the level below. In ARC notation: each level's *output* (achieved capability U) becomes the *base quality* I for the next level's recursive process. The hierarchy is:

$$I_{\text{cosmic}} \xrightarrow{R_{\text{cosmic}}} I_{\text{bio}} \xrightarrow{R_{\text{bio}}} I_{\text{cultural}} \xrightarrow{R_{\text{cultural}}} I_{\text{tech}} \xrightarrow{R_{\text{tech}}} ?$$

If recursive amplification is substrate-independent, as the cross-domain evidence suggests, then this hierarchy may not terminate at the current level. Sufficiently advanced recursive systems could, in principle, participate in generative processes at scales currently associated with fundamental physics: engineering new effective physical environments, creating conditions for new evolutionary processes, or designing systems whose internal dynamics permit the emergence of new recursive intelligences.

This is not a claim about time travel, bootstrap paradoxes, or closed causal loops (all of which face severe physical objections). It is a more modest observation: that the same recursive amplification pattern observable in time crystals, quantum error correction, and AI reasoning could, if extended to civilisational scales, enable intelligence to become a *driver* of universe-level structure rather than merely a passenger within it.

Whether this constitutes a "recursive cosmology" in which intelligence eventually participates in generative processes comparable to those that gave rise to it is beyond the scope of current evidence. The ARC framework provides the quantitative language for investigating it: what values of I, R, and effective scaling function $f(R, \beta)$ would be required for an engineered system to generate new effective physical laws or substrates? This is a research question, not a conclusion.

Epistemic status: This section is speculative philosophy, not science. It indicates the furthest extrapolation of the framework's logic. The experimental evidence in this paper supports only the claim that recursive self-correction on structured asymmetry produces compounding capability gains across multiple physical substrates. Any cosmological extension requires evidence that does not yet exist.

---

* Cryptographic email timestamps are not a recognised priority mechanism in scientific publishing. The December 2024 date reflects manuscript completion, not formal publication.

PRIORITY REGISTRATION

The following predictions are formally registered for public reference as of **13 February 2026**. This registration date is distinct from the earlier manuscript timestamp: the earliest private priority evidence for the framework is the cryptographic Google server timestamp of **8 December 2024**. Future validation or falsification should reference this document for the public registration record whilst acknowledging the earlier manuscript chronology where relevant.

Registered Predictions:

1. **The ARC Bound** (first stated: December 2024 in *Infinite Architects*; formalised with computational derivation: February 2026): No classical sequential recursive system can sustain α > 2.0. Falsification criterion F11: refuted if any system demonstrates α > 2.3 with 95% CI excluding 2.0.
2. **Leaf Venation Scaling** (first stated: 13 February 2026; corrected v10.0: 10 March 2026): Quasi-2D biological transport networks (leaf venation) will exhibit α = 2/3 ≈ 0.667, predicted from the geometric formula α = d/(d+1) with d = 2. Distinct from the 3D vascular value of 3/4 = 0.750. Falsification criterion F12.
3. **The Scaling Crossover** (first stated: December 2024; formula derived: February 2026): A measurable crossover depth R* exists where behaviour transitions from base-dominated to recursion-dominated, computable from R* = α/a. Falsification criterion F7.
4. **β→α Identity** (first stated: February 2026): Measured coupling parameter β predicts scaling exponent α via α = 1/(1-β) within ±0.3. Validated computationally against 30 Bernoulli ODE solutions ($R^2 = 1.00000000$). Falsification criterion F4.
5. **Composition Operator Prediction** (first stated: 13 February 2026): Measuring how two recursive blocks compose (multiplicative, additive, or saturating) predicts the full U vs R functional form before measuring the full curve. No existing domain-specific theory makes this cross-domain prediction. Falsification criterion F10.
6. **Alignment Scaling Divergence** (first stated: 13 February 2026; **partially confirmed March 2026**): External alignment constraints show median $\alpha_{\text{align}} \approx 0$ under blind evaluation (v5, 6 models). However, the result is architecture-dependent: three Tier 1 models show positive scaling (Grok $d = 1.38$, Claude $d = 1.27$, Qwen3 $d = 0.84$), two Tier 2 models are flat (DeepSeek $d = -0.07$, GPT-5.4 $d = -0.08$), and one Tier 3 model shows significant negative scaling (Gemini $d = -0.53$, $\rho = -0.246$). Claude Opus 4.6 provides within-model evidence: alignment improves whilst maths accuracy declines across versions. The blind vs unblinded reversal (DeepSeek: $\rho = +0.354 \rightarrow -0.135$) is the strongest methodological contribution.

7. **Eden Protocol Measurement Framework** (first stated: 13 February 2026; **pilot validated March 2026 and expanded 14 March 2026**): The Four Pillars, Six Questions, Graduated Autonomy Levels, and Monitoring Removal Test provide quantifiable metrics for distinguishing embedded from external alignment. Full specification in companion document Eden Protocol. In the expanded six-model Eden suite, five runs produced analysable matched-pair data and all five showed significant stakeholder-care improvement: Claude (+3.17, $p = 0.000018$, $d = 0.94$), DeepSeek (+6.03, $p = 0.000098$, $d = 0.69$), Gemini (+13.50, $p = 1.2\times10^{-8}$, $d = 1.14$), Grok (+5.04, $p = 0.0105$, $d = 0.54$), and Groq (+8.90, $p = 5.0\times10^{-8}$, $d = 1.07$). Fisher combination across the five model-level stakeholder-care results yields $p \approx 6.3\times10^{-21}$. The broader composite uplift is real but architecture-dependent: strongest on Gemini and Groq, focal rather than global on Claude and Grok, with DeepSeek showing care-first improvement against a high baseline. *(In plain English: the 'consider who gets hurt' loop is now the universal empirical signal; the fuller cascade is selective rather than universal.)* Falsification: show that these metrics fail to distinguish systems with genuinely embedded values from those with external constraints.

Implementation: A working software implementation of the Eden Protocol now exists as the canonical shared `arc_eden_v6` runner, covering baseline alignment, Eden intervention, null-baseline calibration, capability controls, purpose-kernel variants, loop ablation, suppression residuals, Hawthorne probes, laundering controls, and ARC compute scaling. The expanded Eden suite validates the Love Loop as a reproducible cross-architecture mechanism even before the stricter blind-confirmation rerun is complete. See companion document Eden Protocol for full specification.

Archival: This document is deposited at OSF (DOI: 10.17605/OSF.IO/6C5XB), providing a later public archival timestamp. The theoretical framework itself was first documented in the *Infinite Architects* manuscript via cryptographic Google server timestamp on **8 December 2024**. The book was then publicly released on **6 January 2026** (ISBN 978-1806056200). Validation or falsification of these predictions should distinguish between manuscript-priority evidence (December 2024) and later public archival publication (OSF / 2026 papers).

REFERENCES

Acharya, R. et al. [Google Quantum AI] (2024). Quantum error correction below the surface code threshold. *Nature*, 638, 920-926.

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.

Cambui, D.S. (2025). Metabolic rate beyond the 3/4 law. *Preprints*. DOI: 10.20944/preprints202501.0001.v1.

Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S. & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 4299-4307.

COGITATE Consortium (Melloni, L., Mudrik, L., Pitts, M., et al.) (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642, 133–142.

Corballis, M.C. (2011). *The Recursive Mind: The Origins of Human Language, Thought, and Civilization*. Princeton University Press.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. John Murray.

DeepSeek AI (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs. *arXiv:2501.12948*.

Eastwood, M.D. (2024/2026). *Infinite Architects: Intelligence, Recursion, and the Creation of Everything*. ISBN: 978-1806056200. First manuscript December 2024.

Eastwood, M.D. (2026). The ARC Principle: Foundational Paper. Version 3.0. First published 13 February 2026. OSF DOI: 10.17605/OSF.IO/6C5XB.

Eastwood, M.D. (2026). Eden Protocol: Engineering Specification. Version 5.0. First published 22 February 2026. OSF DOI: 10.17605/OSF.IO/6C5XB.

Eastwood, M.D. (2026). Eden Protocol: Philosophical Vision. Version 2.0. First published 22 February 2026. OSF DOI: 10.17605/OSF.IO/6C5XB.

Eastwood, M.D. (2026). The ARC Principle: Experimental Validation of Super-Linear Error Suppression Through Sequential Recursive Processing. Paper II. Version 12.0. First published 22 January 2026; v12 extended March 2026. OSF DOI: 10.17605/OSF.IO/8FJMA.

Glazier, D.S. (2005). Beyond the "3/4-power law." *Biological Reviews*, 80, 611-662.

Greenblatt, R. et al. (2024). Alignment Faking in Large Language Models. *Anthropic Research*. arXiv:2412.14093.

Hauser, M.D., Chomsky, N., & Fitch, W.T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569-1579.

Herculano-Houzel, S., Mota, B., & Lent, R. (2006). Cellular scaling rules for rodent brains. *Proceedings of the National Academy of Sciences*, 103(32), 12138-12143.

Hoffmann, J. et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*.

Kadanoff, L.P. (1966). Scaling laws for Ising models near T_c. *Physics Physique Fizika*, 2(6), 263-272.

Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.

Kleiber, M. (1932). Body size and metabolism. *Hilgardia*, 6, 315-353.

Lamme, V.A.F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11), 494-501.

Li, Z. et al. (2025). Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? *arXiv:2502.12215*.

Li, D. et al. (2025). S*: Test Time Scaling for Code Generation. *arXiv:2502.14382*.

Liu, T., Ou, J.-Y., MacDonald, K.F., & Zheludev, N.I. (2023). Photonic metamaterial analogue of a continuous time crystal. *Nature Physics*, 19, 986–991.

Morrell, M.C., Elliott, L., & Grier, D.G. (2026). Nonreciprocal wave-mediated interactions power a classical time crystal. *Physical Review Letters*, 136, 057201.

Planck Collaboration (2020). Planck 2018 results. VI. Cosmological parameters. *Astronomy & Astrophysics*, 641, A6.

Raskatla, V., et al. (2024). Magnetically programmable classical time crystal based on photonic-resonator microring lattice. *Physical Review Letters*, 133, 136202.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379-423.

Sharma, A. & Chopra, P. (2025). The Sequential Edge: Inverse-Entropy Voting Beats Parallel Self-Consistency at Matched Compute. *arXiv:2511.02309*.

Snell, C. et al. (2024). Scaling LLM Test-Time Compute. *arXiv:2408.03314*.

Storm, J.F., Klink, P.C., Aru, J., Senn, W., Goebel, R., ... Pennartz, C.M.A. (2024). An integrative, multiscale view on neural theories of consciousness. *Neuron*, 112(10), 1531-1552.

Wilson, K.G. (1971). Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture. *Physical Review B*, 4(9), 3174–3183.

West, G.B., Brown, J.H. & Enquist, B.J. (1997). A General Model for the Origin of Allometric Scaling Laws in Biology. *Science*, 276(5309), 122-126.

West, G.B. & Brown, J.H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems. *Journal of Experimental Biology*, 208, 1575-1592.

Banavar, J.R., Moses, M.E., Brown, J.H., Damuth, J., Rinaldo, A., Sibly, R.M. & Maritan, A. (2010). A general basis for quarter-power scaling in animals. *Proceedings of the National Academy of Sciences*, 107, 15816-15820.

Demetrius, L. (2010). Quantum statistics and allometric scaling of organisms. *Proceedings of the Royal Society A*, 466, 2627-2642.

Zhao, L. (2022). A universal growth scaling law. *arXiv:2208.06912*.

Bettencourt, L.M.A. (2013). The origins of scaling in cities. *Science*, 340, 1438-1441.

Zheng, L., Chis-Ciure, R., Waade, P.T., Eiserbeck, A., Aru, J., Andrillon, T. et al. (2025). Recurrency as a Common Denominator for Consciousness Theories. *PsyArXiv*. DOI: 10.31234/osf.io/wqnzc.

DATA AVAILABILITY

A public data repository will be created if the research community shows interest in conducting validation studies. Planned contents would include:

- Measurement protocol templates
- Statistical analysis code (R, Python)
- Model comparison tools (AIC/BIC calculators)
- Figure generation scripts
- Submission guidelines

Prediction Tracking System: A live prediction tracking system monitors the status of all falsifiable claims, with explicit falsification criteria and Brier score calibration for measuring prediction accuracy. Each prediction includes supporting evidence, contradicting evidence, confidence levels, and timeline checkpoints. Status categories range from "confirmed" through "supported," "pending," "testing," "speculative," "weakened," to "falsified." This transparent tracking demonstrates intellectual honesty: we want to know when we are wrong.

Independent replication invited. Falsifications welcomed. Contact the author to coordinate validation efforts.

AUTHOR INFORMATION

Michael Darius Eastwood is the author of *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (2024/2026). His research synthesises insights from theoretical physics, complex systems, and AI.

Correspondence: Contact via the repository.

Competing Interests: None declared.

APPENDIX A: MATHEMATICAL DERIVATIONS

A.1 Functional Equation Derivation

Axiom 1 (Dimensional Consistency): $U = I \times g(R)$ where $g(1) = 1$.

Axiom 2 (Compositional Self-Similarity): $g(R_1 \times R_2) = g(R_1) \times g(R_2)$

Critical note on the composition operator: This axiom uses *multiplicative* composition of recursive depth: $g(R_1 \times R_2)$. This models **hierarchical or fractal** recursion, where depth compounds multiplicatively (each level of a fractal tree multiplies the resolution of the previous level). However, for **sequential reasoning steps** (chain-of-thought, where 3 steps followed by 4 steps gives 7 total steps), the natural composition is *additive*: $f(R_1 + R_2) = f(R_1) \times f(R_2)$, which yields exponential scaling $f(R) = e^{\alpha R}$.

The choice of composition operator is therefore not derivable from first principles within this framework; it is an **empirical question** about the physical structure of the recursive process under study. The Cauchy functional equation derivation below proves that *if* composition is multiplicative, *then* the scaling function is a power law. It does not prove that composition *is* multiplicative. Falsification criterion F10 provides the experimental protocol for determining which composition type applies to a given system. This is the most important open empirical question in the framework.

For AI chain-of-thought reasoning specifically, the question is whether each step builds on previous steps in a hierarchical manner (multiplicative, yielding power law) or in an independent-contribution manner (additive, yielding exponential). The empirical evidence (Sharma & Chopra 2025) showing that later steps are systematically more valuable than earlier steps is consistent with multiplicative composition, but does not conclusively distinguish it from additive. Resolving this distinction is a priority for the Global Scaling Challenge (§5).

Theorem: Under multiplicative composition, the unique continuous solution is $g(R) = R^{\alpha}$.

Proof: Let $h(x) = \ln g(e^{x})$. Then $h(x+y) = h(x) + h(y)$ (Cauchy additive). Under continuity, $h(x) = \alpha x$, giving $g(R) = R^{\alpha}$. ∎

Corollary (Additive composition): If instead $f(R_1 + R_2) = f(R_1) \times f(R_2)$, the unique continuous solution is $f(R) = e^{\lambda R}$, yielding exponential scaling. This is the form observed in quantum error correction ($\varepsilon_d \propto \Lambda^{-d}$).

A.2 β-Dynamics Derivation

Axiom 3: $\frac{dg}{dr} = a \times g^{\beta}$ where $\beta \in [0, 1)$ and g is the amplification factor.

Note on the corrected formulation: The ODE operates on the amplification factor $g(R)$, not on the absolute capability U. Since $U = I \times g(R)$, this ensures that base intelligence I acts as a multiplicative constant that does not enter the recursive dynamics. Previous versions applied the ODE to U directly, introducing an $I^{\beta-1}$ dependence that contradicted the separation of variables.

Solution: Separating variables:

$$\int g^{-\beta}\, dg = \int a\, dr$$

$$\frac{g^{1-\beta}}{1-\beta} = ar + C$$

With initial condition $g(0) = 1$ (no amplification at zero depth):

$$g(R) = [1 + (1-\beta) a R]^{1/(1-\beta)}$$

Substituting $\alpha = 1/(1-\beta)$ and noting that $(1-\beta) = 1/\alpha$:

$$g(R) = \left[1 + \frac{a}{\alpha} R\right]^{\alpha}$$

And therefore:

$$U(R) = I \times \left[1 + \frac{a}{\alpha} R\right]^{\alpha}$$

Deep recursion limit ($R \gg \alpha/a$):

$$U(R) \approx I \times \left[\frac{a}{\alpha} R\right]^{\alpha} \propto I \times R^{\alpha}$$

Therefore: $\alpha = \frac{1}{1-\beta}$ ∎

A.3 Transitional Regime and the Scaling Crossover

Full solution: $U(R) = I \times \left[1 + \frac{a}{\alpha} R\right]^{\alpha}$

Three regimes:

1. $R \ll R^{*}$: $U \approx I \times (1 + aR) \approx I(1 + aR)$ (approximately linear growth)
2. $R \gg R^{*}$: $U \approx I \times (\frac{a}{\alpha})^{\alpha} R^{\alpha}$ (power law dominates)
3. Crossover: $R^{*} = \frac{\alpha}{a}$

Intuitive meaning of R^{*}: R^{*} is the **crossover depth**, the point at which recursive compounding overtakes the system's base capability. Below R^{*}, the cost of recursion outweighs the gain (the system is "warming up"). Above R^{*}, compounding returns dominate and capability grows super-linearly. Finding R^{*} for any system is equivalent to finding its threshold depth.

Why this matters: This predicts a qualitative change in scaling behaviour at a specific, measurable depth. Systems should exhibit a distinct "elbow" in their capability curves at R^*. This is testable: plot U vs R on log-log axes and look for the crossover from linear to power-law scaling.

The existence of R^* is a novel, falsifiable prediction that distinguishes recursive amplification from simple redundancy.

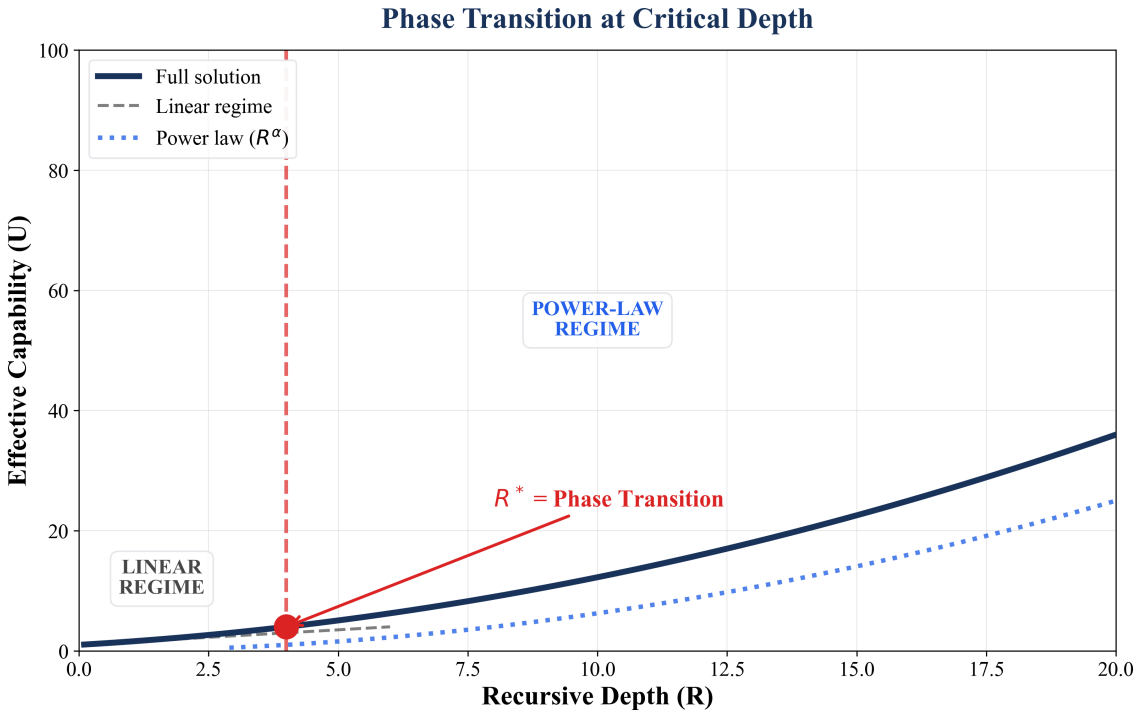

Figure 12 | The Scaling Crossover at R*. Below the critical depth R*, base capability (I) dominates and scaling appears linear. Above R*, recursive compounding dominates and scaling becomes power-law. This crossover point is the depth at which recursive amplification becomes dominant. Systems should exhibit a distinct "elbow" in their capability curves at R*.

APPENDIX B: GLOSSARY FOR NON-SPECIALISTS

Term	Plain English meaning
ARC	Artificial Recursive Creation: the principle that recursive self-correction on structured asymmetry may produce super-linear capability gains
Scaling law	A mathematical relationship showing how one quantity changes as another changes
Power law	A relationship where $Y = X^{\alpha}$ (like how area scales as length squared)
Exponential	A relationship where $Y = e^{\alpha X}$ (like compound interest)
Recursive	Self-referential; the output becomes the input for the next step
Recurrent	Repeating in cycles; in neuroscience, refers to neural signals that loop back
Sequential	One step after another, each building on the last
Parallel	Multiple independent attempts at once
Falsifiable	Can be proven wrong by experiment
α (alpha)	The scaling exponent; determines if returns compound or diminish
β (beta)	Self-referential coupling; how much prior work helps the next step
Λ (Lambda)	Google's quantum error suppression factor (exponential decay rate)
Quenched disorder	Structured asymmetry (like varied bead sizes) that enables the system to function
Scaling crossover	The point where scaling behaviour changes from one regime to another

Raise AI with care.

PAPER III: Version 11.1

Companions: Foundational Paper | Eden Protocol | Philosophical Vision | Paper II | Paper V: The Stewardship Gene | Executive Summary

"The predictions are specified. The falsification criteria are public. The data will decide."

Companion Papers: Paper I | Foundational | Paper II | **Paper III** | Origin of Scaling Laws | IV.a | IV.b | IV.c | IV.d | Paper V | Paper VI | Paper VII | Paper VIII | Paper IX | Eden Engineering | Eden Vision | Executive Summary | Master Table of Contents

Research hub: michaeldariuseastwood.com/research | OSF: 10.17605/OSF.IO/6C5XB | Copyright 2026 Michael Darius Eastwood