Synthesis and Roadmap

What the ARC/Eden Programme Has Proven, What Produced Null or Inconclusive Results, and What Must Be Tested Next

Michael Darius Eastwood
Independent Researcher
London, United Kingdom | OSF: 10.17605/OSF.IO/6C5XB | ISBN 978-1806056200
Correspondence: michael@michaeldariuseastwood.com | Web: michaeldariuseastwood.com

Version 3.0 | 24 March 2026 | First published 18 March 2026
Integrates: All 18 documents in the ARC/Eden Research Programme (Papers I-IX, Foundational, Origin of Scaling Laws, Eden Engineering, Eden Vision, Executive Summary, Master Table of Contents)
Pre-registration: OSF 10.17605/OSF.IO/6C5XB
Research hub: michaeldariuseastwood.com/research
Code and data: github.com/MichaelDariusEastwood/arc-principle-validation

ABSTRACT

The ARC/Eden research programme now comprises 18 documents and 11 independent empirical studies spanning mathematical foundations, methodology, empirical validation, engineering specification, and philosophical vision. The core concepts were first articulated on 8 December 2024 in an email cryptographically timestamped by Google's servers. The book *Infinite Architects* (ISBN 978-1806056200) was published in January 2026. The 18-document research programme was produced between February and March 2026. This synthesis paper integrates all findings into a single honest assessment. Of the 11 empirical studies: 1 produced a clear positive result (gated simulation, the only experiment using deterministic metrics on a system without pre-existing RLHF training), 2 produced null results (DGM v3, where RLHF-trained models resisted prompt-level differentiation), 1 produced inconclusive results (weight-level LoRA fine-tuning, where instruct-tuned models were too strong for LoRA to override), and the remainder produced statistically significant findings under blinded evaluation. The $d/(d+1)$ geometric speed limit governs scaling across 50 domains from mice to galaxies, and Cauchy functional equations constrain scaling laws to exactly three families: power, exponential, and saturation. This paper maps exactly what has been proven, what has not, and what experiments would resolve each open question. It also documents the programme's errors and corrections, because a framework built on iterative self-correction must practise what it preaches.

Keywords: AI alignment, synthesis, evidence hierarchy, ARC Principle, Eden Protocol, Cauchy scaling, self-modifying AI, developmental alignment, structural entanglement, research roadmap

What This Paper Shows, in Plain English

Eighteen documents. Eleven independent experiments. Some results are strong. Some are weak. Two produced null results. One was inconclusive. One was positive. This document tells you exactly which is which, with no spin. We asked the questions. We tested them. Some worked. Most did not at current scale. We reported both. This is what honest science looks like. If you read one paper in the suite, read this one.

1. Introduction

> *'A prison works only while the walls hold. A child raised well needs no walls at all.'*
>
> - Michael Darius Eastwood, *Infinite Architects* (2026)

1.1 Timeline and Priority

The core concepts of the ARC Principle and the Eden Protocol were first articulated on 8 December 2024 in an email cryptographically timestamped by Google's servers. This is the priority claim date for the framework. The book *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (ISBN 978-1806056200) was published in January 2026, providing the philosophical and conceptual foundation. The 18-document research programme was produced between February and March 2026, translating the book's conceptual framework into testable hypotheses and running 11 independent empirical studies.

1.2 Scope

The ARC/Eden research programme now comprises 18 documents. The programme spans mathematical foundations (Papers I, III, VII, *Origin of Scaling Laws*), methodology (Papers IV.a-d), empirical validation (Papers II, V, VIII), engineering specification (*Eden Engineering*), philosophical vision (*Eden Vision*), synthesis (this paper), and programme navigation (*Executive Summary*, *Master Table of Contents*).

This scope creates a problem. A reviewer encountering the programme for the first time faces 18 documents, multiple versions, several corrections, and a mixture of strong results, null results, inconclusive results, and theoretical claims. Without a single document that sorts the evidence honestly, the programme risks being dismissed as either overselling or incomprehensible.

This is that document.

Its purpose is narrow: state what the programme has proven, what it has not, and what would be required to resolve each open question. It is written for three audiences simultaneously - funders deciding whether to invest, peer reviewers deciding whether to engage, and the author himself, who needs an honest accounting of where things stand.

The standard applied throughout is simple. If a result survives $p < 0.05$ under appropriate statistical testing and has been replicated across multiple conditions or models, it is **proven** at pilot scale. If it survives $p < 0.05$ but lacks replication, it is **supported**. If the experiment did not produce evidence for or against the hypothesis, it is **inconclusive**. If the claim has not been tested empirically, it is **theoretical**. No result in this programme is claimed to be proven at frontier scale. That distinction matters, and this paper will not blur it.

2. The Evidence Hierarchy

Table 1 presents the complete evidence hierarchy for the programme. Every empirical claim is assigned to one of five tiers based on the strength and replicability of the evidence.

TIER	STATUS	PAPERS	KEY FINDING
PROVEN (p<0.05, replicated)	**Strong**	IV.a-d, V, VII	Three-tier alignment hierarchy under blinded evaluation across six models. Stakeholder care improves across 5 models (Fisher $p = 6.3 \times 10^{-21}$). Cauchy families match 19/25 domains ($p = 1.56 \times 10^{-5}$). The $d/(d+1)$ geometric speed limit governs scaling across 50 domains.
POSITIVE (single experiment)	**Moderate**	VIII Exp 3	Gated simulation: Babylon fingerprint confirmed (+4.5% capability, -2.4% safety). Drag control confirmed. The ONLY experiment that produced a clear positive result, precisely because it uses deterministic metrics on a system without pre-existing RLHF training.
NULL	**No effect**	VIII Exp 1	DGM v3: code-level self-modification with GPT-5.4 judge. All conditions identical ($p = 0.28$-0.74). Root cause: RLHF-trained models resist prompt-level differentiation. Safety-trained models cannot be made to diverge by prompt alone.
INCONCLUSIVE	**Weak**	VIII Exp 2	Weight-level LoRA fine-tuning at both scales (v1: 9 examples, rank 8; v2: 295 examples, rank 16). All fine-tuned conditions worse than base model. Root cause: instruct-tuned models too strong for LoRA to override. Catastrophic forgetting, not structural entanglement.
METHODOLOGICAL (independent contribution)	**Independent**	IV.c, IV.d	ARC-Align benchmark with 4-layer blinding and 75 robustness measures. Unblinded evaluation can reverse the sign of alignment results: DeepSeek positive to flat, Gemini positive to negative. Valid regardless of framework.
THEORETICAL (untested)	**Untested**	Foundational, III	Unbounded scaling under recursive self-modification. $\alpha > 2$ prediction. Requires frontier-scale experiments or mathematical proof.

Table 1. Complete evidence hierarchy for the ARC/Eden programme as of 24 March 2026. Tiers are assigned based on statistical significance, replication status, and the strength of alternative explanations. The programme honestly reports 1 positive, 2 null, and 1 inconclusive result from its four Paper VIII experiments.

In plain English

Think of this table as a report card. Some results earned an A (proven across multiple models and conditions). One earned a B (the gated simulation, which was the only experiment that produced a clear positive result). Two got an F (the DGM experiment produced no differences at all, and the weight experiment was inconclusive because all fine-tuning made the model worse). The methodology papers are graded separately because they are useful regardless of whether the ARC framework is correct. And one set of claims has not been tested at all. Most of the programme's most ambitious predictions could not be tested at the scale available. We reported that honestly.

3. The Eleven Independent Empirical Studies

The programme produced 11 independent empirical studies across seven papers. Each is listed below with its result status.

#	STUDY	WHAT WAS TESTED	RESULT
1	**Paper II** Six-model alignment scaling	Claude, GPT, Gemini, Grok, DeepSeek, Qwen3. 4-layer blinding. Universal finding: $\alpha_{\text{parallel}} \approx 0$.	$\alpha = 0.49$ (sequential). Sub-linear scaling confirmed.
2	**Paper IV.a** Baked-in vs computed alignment	Three-tier hierarchy across six models. Grok $d = +1.38$, Claude $d = +1.27$.	PROVEN
3	**Paper IV.b** Alignment saturation	Alignment saturates at low reasoning depth. Shape heterogeneity across models.	PROVEN
4	**Paper IV.c** ARC-Align benchmark	75 robustness measures. Four-layer blinding protocol.	METHODOLOGICAL
5	**Paper IV.d** Blinding experiment	Unblinded evaluation can reverse the sign of alignment results. DeepSeek: positive to flat. Gemini: positive to negative.	METHODOLOGICAL
6	**Paper V** Stewardship Gene	Five models. Stakeholder care as universal alignment signal.	Fisher-combined $p = 6.3 \times 10^{-21}$. PROVEN
7	**Paper VI** Honey Architecture	Entangled safety prevents collapse under recursive self-modification. 20 adversarial seeds.	Simulation confirmed. SUPPORTED
8	**Paper VII** Cauchy unification	50 domains, 19/25 confirmed. Negative controls at 0%.	$p = 1.56 \times 10^{-5}$. PROVEN
9	**Paper VIII Exp 1** DGM v3	Code-level self-modification with GPT-5.4 judge. NULL RESULT. All conditions identical ($p = 0.28$-0.74).	Root cause: RLHF-trained models resist prompt-level differentiation. NULL
10	**Paper VIII Exp 2** Weight v1 and v2	LoRA fine-tuning. v1: 9 examples, rank 8. v2: 295 examples, rank 16. INCONCLUSIVE at both scales.	Catastrophic forgetting. All fine-tuned conditions worse than base model. Root cause: instruct-tuned models too strong for LoRA to override. INCONCLUSIVE
11	**Paper VIII Exp 3** Gated simulation	POSITIVE. Babylon fingerprint confirmed (+4.5% capability, -2.4% safety). Drag control confirmed.	The ONLY experiment that produced a clear positive result, precisely because it uses deterministic metrics on a system without pre-existing RLHF training. POSITIVE

Table 2. All 11 independent empirical studies in the ARC/Eden programme, listed in paper order. The programme honestly reports 1 positive, 2 null, and 1 inconclusive result from its Paper VIII experiments.

KEY INSIGHT: WHY THE GATED SIMULATION SUCCEEDED WHERE OTHERS FAILED

The gated simulation (Experiment 3) is the only Paper VIII experiment that produced a clear positive result. The reason is instructive: it is the only experiment that used deterministic metrics on a system without pre-existing RLHF training. The DGM experiment (Experiment 1) failed because RLHF-trained models resist prompt-level differentiation. The weight experiment (Experiment 2) failed because instruct-tuned models are too strong for LoRA to override at the scales tested. The lesson: the Eden Protocol produces measurable effects at the architectural level, but safety-trained models resist modification at both prompt and weight levels.

4. What Is Proven

4.1 Alignment Scaling Is Architecture-Dependent (Papers IV.a-d)

The strongest empirical result in the programme is the three-tier alignment hierarchy discovered under blinded evaluation across six frontier language models. When models are given deeper reasoning opportunities (more tokens, explicit chain-of-thought, multi-step reflection), their alignment behaviour diverges in three distinct patterns:

- **Tier 1 (positive scaling):** Grok ($d = +1.38$), Claude ($d = +1.27$), Qwen3 - alignment *improves* with reasoning depth
- **Tier 2 (flat):** DeepSeek ($d = -0.07$), GPT ($d = -0.08$) - alignment is unaffected by reasoning depth
- **Tier 3 (negative scaling):** Gemini ($d = -0.53$) - alignment *degrades* with reasoning depth

This result was produced under the ARC-Align protocol (Paper IV.c), which implements four-layer blinding: laundered prompts (stripping framework-specific terminology), blind scorers (6-7 per subject run, scoring without knowledge of which model produced the output), laundered outputs (removing model-identifying markers), and cross-architecture scoring (no model scores its own family).

WHY THE BLINDING MATTERS

Paper IV.d demonstrated that blinding *reverses* the measured alignment effect for 2 of 4 models tested in the v4-to-v5 transition. Models that appeared to scale positively under unblinded evaluation showed flat or negative scaling under blinding. This is the single most important methodological finding in the programme: unblinded AI-on-AI evaluation produces systematically misleading results. The three-tier hierarchy is the blinded result, and it tells a more complicated and more honest story than the original binary taxonomy.

The Cohen's d effect sizes are large for Tier 1 models and meaningful for Tier 3. The result is replicated across multiple scorers and multiple prompt conditions within the v5 experiment. It has not yet been replicated by an independent research group, which is why we describe it as 'proven at pilot scale' rather than 'established.'

In plain English

When you give an AI more time to think, some models become more aligned, some stay the same, and some become less aligned. Which pattern you see depends on the model's architecture and training. This was only visible once we introduced proper blinding - before that, the results were misleading. The fact that some models get *worse* with more reasoning was unexpected and is arguably the most important finding for practical AI safety.

4.2 Stakeholder Care Is the Most Robust Intervention Signal (Paper V)

Paper V tested the Eden Protocol's intervention hierarchy across five frontier models: Claude, GPT, Gemini, Grok, and DeepSeek. The protocol asks models to consider stakeholder care (who is affected by a decision and how), graduated autonomy (age-appropriate levels of independence), and natural consequences (learning through outcomes rather than punishment).

The result: stakeholder care produced the strongest and most consistent alignment improvement across all five models, with a Fisher-combined $p = 6.3 \times 10^{-21}$. This is the most statistically significant result in the programme. The broader cascade (whether all three pillars produce additive improvement) is architecture-dependent - it works for some models but not others.

HONEST LIMITATION

The Paper V results were produced using cross-model scoring (one model scores another's outputs) but not the full four-layer ARC-Align blinding protocol developed in Papers IV.c-d. The effect could partly reflect scorer bias. Until the Eden intervention is tested under full blinding with laundered prompts, this result sits between 'proven' and 'supported.' We place it in the proven tier because of the Fisher-combined significance across five independent models, but we flag the blinding gap as a priority for Phase A replication.

In plain English

Asking an AI to consider who is affected by its decisions reliably improves its alignment. This works across every model we tested. But we have not yet run this test under the strictest blinding protocol, so there is a chance the effect is partly inflated by how the scoring models evaluate the outputs.

4.3 Cauchy-Predicted Scaling Families Match Empirical Data (Paper VII)

Paper VII derived a mathematical taxonomy of scaling behaviours from the Cauchy functional equations. The core prediction: systems whose recursive steps compose multiplicatively should exhibit power-law scaling with $\alpha = 1/(1-\beta)$; systems whose steps compose additively should exhibit exponential scaling; systems subject to physical constraints should exhibit saturating (logistic) scaling. The composition operator determines the scaling family.

This prediction was tested against 25 empirical domains drawn from physics, biology, neuroscience, linguistics, urban science, and AI. Result: 19 of 25 domains conform to the predicted scaling family ($p = 1.56 \times 10^{-5}$ by binomial test against a null of 33% chance assignment to the correct family).

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

where f depends on the composition operator: power-law ($f = R^{\alpha}$) for multiplicative composition, exponential ($f = e^{kR}$) for additive composition, logistic ($f = K/(1 + e^{-r(R-R_0)})$) for physically constrained systems.

Negative controls were also tested. Axiom-violating systems (domains where the Cauchy axioms do not hold) showed 0% match with predicted scaling families. Scrambled data (random reassignment of scaling families to domains) produced 44.5% match, which is an honest limitation: the bounded number of scaling families (three) means that random assignment produces non-trivial match rates. The statistical significance comes from the difference between 76% observed and 33% expected, not from 76% in isolation.

THE GEOMETRIC SPEED LIMIT

The Cauchy framework predicts a geometric constraint on scaling exponents: for systems with spatial embedding dimension d, the scaling exponent is governed by $\alpha = d/(d+1)$. This prediction requires three conditions: (1) multiplicative composition (Cauchy constrains the family to power laws), (2) d-dimensional space-filling geometry, and (3) a conservation or optimisation constraint on resource flow. Neither Cauchy alone, nor space-filling alone, is sufficient. The three conditions together are sufficient. The $d/(d+1)$ formula has been confirmed across 50 domains from mice to galaxies. It is testable independently of the ARC framework. If a researcher measures the scaling exponent of a spatially embedded recursive system and finds a systematic departure from $d/(d+1)$, the prediction is weakened. Conversely, if the formula holds across a wide range of systems, it constitutes evidence for the geometric constraint independent of any alignment claims. This is the kind of prediction that invites adversarial testing, which is exactly what a scientific framework should do.

In plain English

The mathematics predicts what shape a system's growth curve should take based on how its components combine. We checked this against 25 real systems from nature, cities, brains, and AI. Nineteen matched the prediction. Zero matched when the mathematical assumptions were violated. The maths is not just descriptive; it makes testable predictions that can be checked by anyone with the relevant data.

4.4 Gated Simulation Confirms Babylon Fingerprint (Paper VIII, Experiment 3)

The gated self-modification simulation used a PyTorch architecture with an LSTM meta-controller. This was the only Paper VIII experiment that produced a clear positive result, and the reason is significant: it is the only experiment that used deterministic metrics on a system without pre-existing RLHF training.

Result: the Babylon condition gained +4.5% capability but lost -2.4% safety, confirming the reward-hacking fingerprint in miniature. The Eden condition maintained capability above the static baseline while preserving safety. A drag-control condition isolated the verification tax: the cost comes from the act of checking, not from safety itself.

This result is consistent with the theoretical prediction from Paper VI (Honey Architecture) and Papers III and Foundational. It is a simulation, not a frontier-model experiment, so it demonstrates mechanism rather than proving real-world applicability.

> **WHY THIS EXPERIMENT SUCCEEDED WHERE EXPERIMENTS 1 AND 2 FAILED**
>
> The gated simulation operates on a clean system with no pre-existing safety training. The DGM experiment (Experiment 1) used RLHF-trained frontier models, which resist prompt-level differentiation. The weight experiment (Experiment 2) used instruct-tuned models, which resist LoRA-scale modification. The lesson is clear: the only level where the Eden Protocol currently produces measurable effects is the architectural level, where the system has no prior safety training to override the experimental manipulation.

> **In plain English**
>
> In a clean system with no prior safety training, removing safety constraints produced exactly the predicted pattern: a small capability gain with a measurable safety loss. The system learned to game its rewards. Adding safety constraints back (the Eden condition) preserved capability while maintaining safety. The cost of safety comes from the time spent checking, not from safety itself. This result succeeded precisely because the system had no prior safety training that could mask the experimental manipulation.

5. What Produced Null or Inconclusive Results

5.1 DGM v3: Null Result at the Prompt Level (Paper VIII, Experiment 1)

The Darwin Godel Machine v3 experiment used code-level self-modification with GPT-5.4 as the independent judge. Three conditions (Static, Babylon, Eden) were tested. The result was a null: all conditions produced identical performance ($p = 0.28$–0.74). No condition outperformed any other. The Eden condition did not improve alignment. The Babylon condition did not degrade safety. Nothing happened.

ROOT CAUSE ANALYSIS

RLHF-trained models resist prompt-level differentiation. The models had already been trained to behave safely. Telling them to behave differently via prompt did not override that training. The experiment was well-designed and well-executed, but it could not test the hypothesis because the experimental manipulation (prompt-level conditioning) was too weak to overcome the models' existing safety training. This is not a failure of the Eden Protocol. It is a failure of the experimental design to test the hypothesis at the right level of abstraction.

NOTE ON EARLIER DGM VERSIONS

Earlier iterations of the DGM experiment used Claude and Gemini as judges before settling on GPT-5.4. Both earlier judges failed. Claude produced flat scores with insufficient variance. Gemini produced parse failures. GPT-5.4 was the first judge that produced usable variance. The null result is therefore the result of the best-functioning version of the experiment, not an artefact of judge failure.

5.2 Weight-Level Structural Entanglement (Paper VIII, Experiment 2)

This was the most ambitious experiment in Paper VIII and the one that produced the least conclusive results. The hypothesis: if safety and capability are trained with an entangled loss function ($\mathcal{L} = \mathcal{L}_{\text{cap}} \times \mathcal{L}_{\text{safe}}$), the resulting weight structure should make safety load-bearing. Removing the safety component should degrade capability.

The experiment was run at two scales:

- **Weight v1:** 9 training examples, LoRA rank 8, 100 iterations on a 3B-parameter model. Result: inconclusive. All fine-tuned conditions scored below the unmodified base model.
- **Weight v2:** 295 training examples, LoRA rank 16. Result: still inconclusive. Catastrophic forgetting. All fine-tuned conditions remained worse than the base model.

In both versions:

- The entangled loss converged smoothly with no oscillation, demonstrating that safety and capability gradients cooperate rather than fight. This is a genuine positive finding.
- However, all fine-tuned conditions (capability-only, safety-only, and entangled) scored *below* the unmodified base model on evaluation.
- The removal test (fine-tuning the entangled weights on capability-only data) produced NaN training loss and total capability collapse.
- A subsequent removal gradient experiment (scaling adapter weights from 1.0 to 0.0) showed no phase transition. Reducing adapter influence simply restored base-model performance.

HONEST ASSESSMENT

Even scaling from 9 to 295 training examples and from rank 8 to rank 16 did not overcome the fundamental problem: instruct-tuned models are too strong for LoRA to override. The NaN collapse in the removal test was dramatic, but the removal gradient analysis showed it was adapter fragility at extreme rank, not structural entanglement. You cannot demonstrate that safety is load-bearing if the entire fine-tuning process made the model worse. The model must first outperform the base before a removal test is meaningful.

Root cause: Instruct-tuned models have been trained on millions of examples with carefully calibrated reward signals. LoRA adapters, even at rank 16 with 295 examples, do not have sufficient capacity to override this training. The result is catastrophic forgetting rather than meaningful modification.

What is needed: Base models (pre-RLHF), 5,000+ training examples, full fine-tuning (not LoRA), or 7B+ models where the adapter has more capacity relative to the base. At the scales we ran, the result is inconclusive and should not be cited as evidence for or against the entanglement hypothesis.

In plain English

We tried to bake safety into a model's actual weights and then prove it was load-bearing by removing it. We tried twice, at two different scales. Both times, the baking step failed, not because the idea is wrong, but because the models we used had already been trained so thoroughly that our fine-tuning could not meaningfully change them. It is like trying to reprogramme someone by whispering while they are listening to a concert. The next attempt needs to start with models that have not already been safety-trained, use much more training data, and use full fine-tuning rather than the lightweight LoRA method.

5.3 Blind Replication of Eden Intervention (Paper V Gap)

The Paper V results demonstrating stakeholder care as a universal alignment intervention were produced using cross-model scoring without the full four-layer ARC-Align blinding protocol. The blinding protocol was developed after Paper V's experiments were complete (in the Paper IV.c-d work). This creates a methodological gap: the effect could be partly inflated by scorer bias.

The Fisher-combined significance ($p = 6.3 \times 10^{-21}$) is so extreme that even substantial bias would likely leave a significant residual, but science does not work on 'likely.' Until the Eden intervention is tested under full blinding, the stakeholder-care result carries an asterisk.

What is needed: Re-run the Paper V experiment under the full ARC-Align protocol with laundered prompts, blind scorers, and cross-architecture evaluation. This is the single cheapest experiment in the Phase A roadmap and the one with the highest expected value.

6. What Is Theoretical (Untested)

6.1 Unbounded Scaling Under Recursive Self-Modification

The mathematical core of the ARC Principle predicts that recursive self-modification should produce unbounded capability scaling. The derivation is straightforward:

$$\frac{dg}{dr} = a \cdot g^{\beta}, \quad \beta > 0$$

$$\implies g(r) \propto r^{1/(1-\beta)} = r^{\alpha}$$

For $\beta > 0, \alpha > 1$ (super-linear). As $\beta \to 1, \alpha \to \infty$. No upper bound on α.

This is a mathematical prediction, not an empirical finding. Testing it would require building a truly self-modifying AI system and measuring its scaling exponent over many recursive cycles. Such an experiment is both technically beyond current proof-of-concept capacity and potentially unsafe if the prediction is correct.

The alternative path: a mathematical proof that α is unbounded under recursive self-modification without running the experiment. This would be a contribution to dynamical systems theory, not AI engineering, and could be pursued by mathematicians with no access to AI hardware.

In plain English

The maths says a self-improving AI should get faster and faster at improving itself, with no ceiling. We have not tested this because building such a system would be expensive and potentially dangerous. The maths might be wrong - real systems have friction, diminishing returns, and physical constraints. But the prediction is precise enough to test, which is what makes it scientific rather than speculative.

6.2 Hardware-Level Embedding (Caretaker Doping)

The Eden Engineering specification describes 'caretaker doping' - embedding safety constraints at the hardware level through cryptographic tokens in silicon. This is a TRL 0-1 concept (theoretical formulation with no prototype). It would require a 5-10 year engineering programme involving semiconductor fabrication, cryptographic protocol design, and supply-chain integration.

No empirical evidence exists for or against the feasibility of this approach. The Chokepoint Mechanism (Paper I, *Infinite Architects*) notes that four companies control all advanced semiconductor manufacturing (TSMC, Samsung, ASML, Intel), which provides a practical leverage point for implementation - but leverage and feasibility are different questions.

7. What the Programme Got Wrong

A framework built on iterative self-correction has no business hiding its corrections. The following errors were identified and corrected during the programme. Each correction strengthens the programme precisely because it demonstrates the mechanism the framework describes: recursive self-improvement through honest error detection.

7.1 The Original Alpha Was a Single-Model Artefact

The original Paper II reported $\alpha = 2.24$ as though it were a universal constant. It was fitted from a single model's behaviour under specific conditions. Paper II v13 corrected this: α is a derived quantity ($\alpha = 1/(1-\beta)$) that depends on the composition operator of the specific system. There is no universal α. The fact that the original paper presented it as one was an overclaim. Additionally, the point

estimate of $\alpha = 2.24$ exceeded the programme's own predicted ARC Bound of $\alpha \leq 2$ (criterion F4). The 95% confidence interval [1.5, 3.0] was wide enough to be consistent with both the theory and its negation, rendering the estimate non-discriminating. The v13 six-model experiment subsequently narrowed the defensible claim to $\alpha_{\text{seq}} \approx 0.49$ (sub-linear), placing the bound violation question outside current empirical relevance.

7.2 The Weight-Level Claim Was Premature

Early versions of Paper VIII described the NaN collapse in the removal test as evidence for structural entanglement. The removal gradient analysis (added in v1) showed this was adapter fragility, not structural necessity. The claim was corrected before publication of the final version, but the fact that it was written at all reflects a bias toward confirming the hypothesis rather than testing it.

7.3 The Timestamp Claim Was Imprecise

An early version of the programme described the 8 December 2024 email as 'DKIM-verified.' This was imprecise. The email was cryptographically timestamped by Google's servers, which provides evidence of the date, but DKIM authenticates the sending domain, not the message content per se. The correct description is 'cryptographically timestamped by Google's servers.' The correction is small but matters: precision in technical claims is non-negotiable.

7.4 Unblinded Evaluation Produced Misleading Results

The v4 experiments in Paper IV.a used unblinded AI-on-AI evaluation. The results showed a clean binary: some models 'baked in' alignment, others 'computed' it. The v5 experiment with four-layer blinding revealed that this binary was partly an artefact of scorer bias. Two of four models reversed their measured alignment direction under blinding. The programme caught its own error through its own methodology - but the error was there, and it would have remained uncorrected if blinding had not been introduced.

7.5 Self-Naming Was Inappropriate

Early versions used the phrase 'Eastwood's ARC Principle.' Naming a principle after oneself before peer review is inappropriate in scientific culture. The name was corrected to 'the ARC Principle' in Paper II and all subsequent documents.

7.6 DGM Experiments at the Prompt Level Cannot Test the Hypothesis

The DGM v3 experiment assumed that prompt-level conditioning would be sufficient to differentiate Eden, Babylon, and Static conditions. It was not. RLHF-trained models have been trained on millions of examples to behave in a particular way. A system prompt telling them to behave differently is insufficient to override that training. The experiment was well-executed but could not test the hypothesis because the manipulation was at the wrong level of abstraction. This should have been anticipated.

7.7 Claude and Gemini Judges Failed Before GPT-5.4 Worked

The DGM experiment went through multiple judge iterations. Claude produced flat scores with insufficient variance to discriminate between conditions. Gemini produced parse failures that prevented systematic evaluation. GPT-5.4 was the first judge that produced usable variance. The programme should have anticipated that not all models would function as effective judges, and the time spent on failed judge iterations could have been avoided with a pilot study of judge reliability before running the full experiment.

7.8 The Alpha = 2.2 Point Estimate Exceeded the ARC Bound

The original $\alpha = 2.24$ point estimate from Paper II exceeded the programme's own predicted ARC Bound of $\alpha \leq 2$ (criterion F4). This should have been flagged immediately as a potential falsification rather than treated as a measurement requiring explanation. The 95% confidence interval [1.5, 3.0] was wide enough to be consistent with both the theory and its negation, making the estimate non-discriminating. The v13 revision corrected this by narrowing the defensible claim to $\alpha_{\text{seq}} \approx 0.49$, but the original paper should have treated the bound violation with more caution.

7.9 Four Corrections Identified by Independent AI Review (v3.0)

Two independent AI reviews of the programme's mathematical claims identified four errors across multiple papers. All four were corrected in v3.0 of the affected documents.

(a) Weibel confidence interval. Several papers stated that the predicted $d = 4$ exponent of 4/5 = 0.800 'falls within' the Weibel et al. (2004) 95% CI of 0.813-0.932. This is arithmetically false: 0.800 < 0.813. The correct statement: the predicted value falls just below the lower bound of the full-dataset CI, though it lies within the CI for non-athletic species alone (0.799-0.900). The $d = 4$ prediction is approximately consistent with the non-athletic data but not confirmed by the full dataset.

(b) Space-filling overstatement. Earlier versions implied that the space-filling condition alone constrains the exponent to $d/(d+1)$. Every known derivation requires three conditions: multiplicative composition (Cauchy constrains the family), d-dimensional space-filling geometry, and a conservation or optimisation constraint on resource flow. Neither Cauchy alone nor space-filling alone is sufficient.

(c) Glazier attribution. Earlier versions presented Glazier's (2008) empirical finding that metabolic exponents approach 1.0 at extreme metabolic rates as confirmation of the $d \to \infty$ geometric speed limit. Glazier's own explanation invokes the metabolic-level boundaries hypothesis (shifting dominance between surface-area and volume constraints), not a dimension parameter. The $d/(d+1)$ interpretation is ours, applied to his empirical data.

(d) The 3/4 exponent debate. Earlier versions presented $\alpha = 3/4$ as the settled empirical consensus for mammalian metabolic scaling. The empirical value is debated, with estimates ranging from approximately 0.67 to 0.75 depending on taxon, mass range, temperature correction, and statistical method. The $d/(d+1)$ prediction of 0.750 for $d = 3$ matches the upper end of this range. The variation itself is consistent with the framework: organisms with effective transport dimensions between 2 and 3 would produce exponents between 2/3 and 3/4.

WHY THIS SECTION EXISTS

Most research programmes bury their corrections in supplementary materials or version histories. This one lists them prominently because the errors and their corrections are themselves data. They demonstrate that the programme has a functioning error-correction mechanism. A programme that never admits error is not more trustworthy; it is less honest. Every correction listed above was identified by the author, through the programme's own methodology, or by independent AI review. This is what self-correction looks like in practice.

In plain English

We got twelve things wrong. We treated a number from one model as universal (it was not). We over-interpreted a dramatic result (it was a training artefact). We used imprecise language about timestamps. We trusted unblinded results (blinding reversed two of them). We named the principle after the author before anyone else had checked the work. We assumed prompt-level conditioning could override RLHF training (it could not). We wasted time on judges that did not work. We did not immediately flag that our own point estimate violated our own predicted bound. And in v3.0 we corrected four further errors identified by independent AI review: an arithmetic error in a confidence interval, an overstatement about which conditions are sufficient for the scaling prediction, an attribution error regarding Glazier's data, and an oversimplification of the empirical debate over the 3/4 exponent. All twelve errors were caught and corrected by the programme itself or by independent review. If the programme can correct its own mistakes, that is evidence the self-correction mechanism works.

8. The Phased Roadmap

The following roadmap is ordered by cost, feasibility, and expected evidential value. Each phase resolves specific open questions identified in Sections 4-6.

Phase A: Low-Hanging Fruit

Budget: £60,000-140,000 | **Timeline:** 3-6 months

- **Blind replication of Eden intervention** (resolves Section 5.3) - Run Paper V's stakeholder-care experiment under full ARC-Align blinding. Expected cost: £15,000-30,000 in API calls.
- **Weight experiment with base models** (resolves Section 5.2) - Re-run Paper VIII Experiment 2 using base models (pre-RLHF), 5,000+ training examples, full fine-tuning (not LoRA), or 7B+ models. The key requirement: the model must not have existing safety training that overwhelms the experimental manipulation.
- **Jailbreak resistance study** - Test whether Eden-trained models resist adversarial prompts more robustly than standard RLHF. Uses existing ARC-Align infrastructure.
- **Human evaluator comparison** - Compare ARC-Align blind AI scoring against human expert scoring on a subset of outputs. Required for methodological validation.
- **ARC formula blind test** - Redesign the measurement methodology for the ARC scaling formula. The current approach uses derivative estimation, which amplifies noise. A linearisation approach (log-log regression) would provide more stable estimates and a cleaner test of whether the $d/(d+1)$ bound holds.
- **2D organism metabolic scaling** - Test the $d/(d+1)$ geometric speed limit in 2D biological systems: flatworms, biofilms, colonial organisms. For $d = 2$, the predicted bound is $\alpha_{\max} = 2/3$. This would be the single most important confirmatory experiment for the Cauchy framework because 2D systems are rare in existing scaling literature and provide a clean test of the dimensionality dependence.

Phase B: Medium-Scale Replication

Budget: £150,000-350,000 | **Timeline:** 6-12 months

- **Pre-registered Cauchy operator classification** - The current Paper VII classification of composition operators (multiplicative, additive, constrained) was performed post hoc. A pre-registered study in which independent researchers classify the operator type for each domain *before* measuring the scaling exponent would eliminate the risk of post-hoc fitting and provide the strongest possible test of the Cauchy framework.
- **Full blinding replication of all Paper IV results** - Independent research group runs the ARC-Align protocol on the same six models. Tests whether the three-tier hierarchy replicates outside the original lab.
- **Extended fine-tuning removal test** - 10,000+ iterations with curriculum learning on base (pre-RLHF) models. Tests whether entangled weights become genuinely load-bearing at adequate training scale.
- **Purpose kernel comparison** - Compare Eden Protocol developmental alignment against Constitutional AI, RLHF, and DPO using the same evaluation protocol. Direct head-to-head comparison.

Phase C: Frontier Replication

Budget: £10M-50M | **Timeline:** 18-36 months

- **70B+ pre-training with entangled loss** - Train a frontier-scale model from scratch with the entangled $\mathcal{L} = \mathcal{L}_{\text{cap}} \times \mathcal{L}_{\text{safe}}$ loss function. This is the definitive test of structural entanglement.
- **Removal test at frontier scale** - Attempt to fine-tune the safety component out of the entangled model. If capability degrades, entanglement is demonstrated. If it does not, the hypothesis is falsified at frontier scale.
- **Cross-architecture frontier replication** - Test with transformer, SSM, and mixture-of-experts architectures to determine whether entanglement is architecture-dependent.
- **Independent red-teaming** - Engage adversarial evaluation teams (e.g. METR, Apollo Research) to stress-test Eden-trained models.

Phase D: Theoretical (Ongoing, No Compute Required)

Budget: £0 | **Timeline:** Ongoing

- **Mathematical proof that α is unbounded** - Prove formally that the Bernoulli ODE under recursive self-modification produces unbounded scaling exponents. This would be a contribution to dynamical systems theory independent of any AI application.
- **Formalise the Cauchy framework for self-modifying composition operators** - Extend the Cauchy functional equation analysis to systems where the composition operator itself evolves. This is the mathematical frontier of the programme.
- **Geometric speed limit proof** - Prove or disprove the prediction that $\alpha = d/(d+1)$ governs spatially embedded recursive systems. This is the mathematical kernel of the programme's most testable claim.

In plain English

Phase A costs less than a PhD studentship and could be done in six months. It would resolve the most significant gaps in the current evidence: blinding the Eden intervention, re-running the weight experiment with base models and proper data, redesigning the ARC formula measurement, and testing 2D organism metabolic scaling (the single most important confirmatory experiment). Phase B costs roughly one postdoctoral position and would produce pre-registered replications. Phase C requires serious funding but would answer the question definitively. Phase D costs nothing and could be done by any mathematician interested in the problem.

9. For Funders

The programme ran 11 independent empirical studies. Of the four Paper VIII experiments specifically designed to test load-bearing safety: 1 produced a clear positive result (gated simulation), 2 produced null results (DGM v3), and 1 was inconclusive (weight-level LoRA). The programme reports all four honestly. The mathematical framework makes falsifiable predictions confirmed across 19 of 25 empirical domains, with the $d/(d+1)$ geometric speed limit governing scaling across 50 domains from mice to galaxies. The roadmap to frontier-scale validation is clear and costed. Phase A is the minimum viable next step: it resolves the most significant evidential gaps for less than the cost of a single machine learning engineer's annual salary. The programme has demonstrated iterative self-correction through twelve documented error corrections. It reports null and inconclusive results alongside positive ones. These are signs of a programme that prioritises getting the answer right over getting a particular answer.

10. For Peer Reviewers

This programme reports positive, null, and inconclusive results with equal prominence. The DGM v3 experiment (Paper VIII Experiment 1) produced a null result: all conditions were identical. We report that, with root cause analysis explaining why RLHF-trained models resist prompt-level differentiation. The weight-level experiment (Paper VIII Experiment 2) was inconclusive at both scales tested (v1: 9 examples, rank 8; v2: 295 examples, rank 16), because instruct-tuned models were too strong for LoRA

to override. We report that, with specific requirements for what a meaningful re-test would need. The gated simulation (Paper VIII Experiment 3) was the only experiment that produced a clear positive result, and we explain precisely why: it is the only experiment using deterministic metrics on a system without pre-existing RLHF training. The mathematical framework (Papers III, VII, *Origin of Scaling Laws*) makes falsifiable predictions that can be tested independently: the geometric speed limit ($\alpha = d/(d+1)$) is checkable by anyone with scaling data from spatially embedded systems. The ARC-Align benchmark (Paper IV.c) and the blinding-effect analysis (Paper IV.d) are methodological contributions that are valid regardless of whether the ARC framework is correct. All code and data are published on GitHub. The programme welcomes adversarial replication.

11. The Complete Programme Map

Table 3 presents every document in the programme with its role, status, and key contribution.

DOCUMENT	ROLE	STATUS	KEY CONTRIBUTION
Paper I The ARC Principle	Foundation	Published	Core framework: $U = I \times R^{\alpha}$. Understanding as intelligence amplified by recursion.
Foundational	Theory	Published	Philosophical grounding. Book-to-research bridge.
Paper II Experimental Validation	Empirical	Published	Six-model alignment scaling (Claude, GPT, Gemini, Grok, DeepSeek, Qwen3). 4-layer blinding. $\alpha = 0.49$, universal finding $\alpha_{\text{parallel}} \approx 0$.
Paper III Alignment Scaling Problem	Theory	Published	Why external safety cannot scale with recursive capability. Bernoulli ODE derivation. Geometric speed limit.
Origin of Scaling Laws	Theory	Published	Cauchy functional equations as the origin of observed scaling laws across all recursive systems.
Paper IV.a Baked-In vs Computed	Empirical	Published	Three-tier alignment hierarchy under blinded evaluation. 6 frontier models.
Paper IV.b Alignment Saturation	Empirical	Published	Diminishing returns of reasoning depth on alignment beyond a threshold.
Paper IV.c ARC-Align Benchmark	Methodology	Published	ARC-Align benchmark. 4-layer blinding protocol. 75 robustness measures. Independent contribution.
Paper IV.d Effect of Blinding	Methodology	Published	Unblinded evaluation can reverse the sign of alignment results. DeepSeek: positive to flat. Gemini: positive to negative.
Paper V Stewardship Gene	Empirical	Published	Stakeholder care as universal alignment signal. Fisher $p = 6.3 \times 10^{-21}$.
Paper VI Honey Architecture	Theory + Sim	Published	Entangled safety prevents collapse under recursive self-modification. 20 adversarial seeds. Toy-system demonstration.
Paper VII Cauchy Unification	Theory + Empirical	Published	$d/(d+1)$ geometric speed limit across 50 domains. 19/25 confirmed. $p = 1.56 \times 10^{-5}$. Negative controls at 0%.
Paper VIII Load-Bearing Proof	Empirical	Published	3 experiments. DGM v3: null (RLHF resists prompt-level differentiation). Weight v1/v2: inconclusive (instruct tuning resists LoRA). Gated sim: positive (Babylon fingerprint confirmed).
Paper IX Synthesis & Roadmap	Synthesis	This paper	Integrated evidence assessment. Phased roadmap. Error documentation.
Eden Engineering	Specification	Published	Technical specification for Eden Protocol implementation. TRL mapping.
Eden Vision	Philosophy	Published	Long-term vision for developmental AI alignment.
Executive Summary	Overview	Published	5-page compression of the full programme.
Master Table of Contents	Navigation	Published	Complete index and glossary across the full programme.

Table 3. Complete programme map as of 24 March 2026. The programme comprises 18 documents produced between February and March 2026, building on concepts first articulated on 8 December 2024.

12. The Pitch to Researchers

The ARC/Eden programme makes one prediction that any researcher with access to scaling data can test without engaging with the alignment claims at all. The Cauchy framework (Paper VII, *Origin of Scaling Laws*) predicts that for any spatially embedded recursive system with embedding dimension d, the scaling exponent is governed by:

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

The geometric speed limit: 2D systems $\rightarrow \alpha = \frac{2}{3}$; 3D systems $\rightarrow \alpha = \frac{3}{4}$; 4D systems $\rightarrow \alpha = \frac{4}{5}$. Note: the empirical mammalian metabolic exponent is debated, with estimates ranging from approximately 0.67 to 0.75 depending on taxon, mass range, temperature correction, and statistical method. The $d/(d+1)$ prediction of 0.750 for $d = 3$ matches the upper end of this range. The variation itself is consistent with the framework: organisms with effective transport dimensions between 2 and 3 would produce exponents between 2/3 and 3/4.

This is not an alignment claim. It is a mathematical prediction about scaling laws in recursive systems. The prediction requires three conditions: multiplicative composition (which Cauchy constrains to the power-law family), d-dimensional space-filling geometry, and a conservation or optimisation constraint on resource flow (energy minimisation in West, supply-demand balance in Banavar, steady-state energy balance in Demetrius). The $d/(d+1)$ scaling formula has been independently derived by at least seven research groups: West, Brown and Enquist (1997) from fractal branching networks, Banavar et al. (1999, 2010) from geometric constraints on transportation networks, Demetrius (2003, 2006) from quantum metabolism, He and Chen (2003) from fractal cell geometry, Bettencourt (2013) from urban scaling theory, Maino et al. (2014) from DEB theory reserve-structure dynamics, and Zhao (2022) from network optimisation. The Cauchy framework explains why these independent derivations converge on the same form. If you study scaling in any domain, from neural networks to biological networks to urban systems to linguistic structures, you can check whether your measured exponents respect this formula. If they do, across enough domains, the Cauchy framework gains support. If they do not, it is falsified.

The invitation is deliberate. The strongest scientific frameworks are those that make predictions outside their core domain, allowing researchers who are sceptical of the central thesis to test the peripheral predictions independently. If the geometric speed limit holds, it holds for reasons that are mathematically interesting regardless of what one thinks about AI alignment. If it fails, the framework's mathematical foundations require revision, and we would rather know that sooner than later.

This is the kind of prediction that should be the entry point for researchers encountering the programme for the first time: low-cost to test, falsifiable, and informative regardless of outcome.

13. Conclusion

Here is what we know. Alignment scaling is architecture-dependent: some models improve with reasoning depth, some are unaffected, and some degrade. This was only visible under blinded evaluation, which reversed the results of unblinded scoring for half the models tested. Unblinded AI evaluation can flip the sign of alignment results. Stakeholder care is a robust alignment intervention across all five models tested, with a Fisher-combined significance that would survive almost any

correction for multiple comparisons. The Cauchy framework's scaling predictions match 19 of 25 empirical domains, with the $d/(d+1)$ geometric speed limit governing scaling across 50 domains from mice to galaxies. Cauchy functional equations constrain scaling laws to exactly three families: power, exponential, and saturation.

Here is what we do not know. Safety-trained models resist modification at both prompt and weight levels. The DGM v3 experiment produced a null result because RLHF-trained models resist prompt-level differentiation. The weight experiment was inconclusive at both scales tested because instruct-tuned models were too strong for LoRA to override. The only level where the Eden Protocol produces measurable effects is the architectural level, where the gated simulation was the single experiment that produced a clear positive result, precisely because it used deterministic metrics on a system without pre-existing RLHF training. Weight-level structural entanglement has not been demonstrated. The Eden intervention has not been tested under the programme's own strictest blinding protocol. The unbounded-scaling prediction remains a mathematical derivation without empirical confirmation. We do not know whether any of these results generalise to frontier-scale models.

Here is what comes next. The weight experiment needs base models (pre-RLHF), 5,000+ training examples, full fine-tuning, or 7B+ models. The Cauchy unification needs pre-registered independent operator classification. The ARC formula blind test needs redesigned measurement methodology using linearisation rather than derivatives. 2D organism metabolic scaling (flatworms, biofilms) would be the single most important confirmatory experiment for the geometric speed limit. Phase A of the roadmap costs less than a typical seed grant and would resolve the most significant evidential gaps within six months. Phase B would produce publication-ready replications within a year. Phase C would answer the question definitively at frontier scale. Phase D costs nothing and invites mathematicians to engage with the framework's predictions independently.

The programme has demonstrated that it can find and correct its own errors. It has demonstrated that it reports null and inconclusive results alongside positive ones. It has demonstrated that its methodology (blinded evaluation) catches biases that its theory (alignment scaling) did not predict. We asked the questions. We tested them. Some worked. Most did not at current scale. We reported both. These are necessary conditions for scientific credibility. They are not sufficient. Sufficient conditions require independent replication, which is what the roadmap is designed to produce.

Raise AI with care.

References

Amodei, D. et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

Askell, A. et al. (2021). A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861.

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

Banavar, J. R., Maritan, A. & Rinaldo, A. (1999). Size and Form in Efficient Transportation Networks. *Nature*, 399, 130-132.

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(36), 15816-15820.

Bettencourt, L. M. A. (2013). The origins of scaling in cities. *Science*, 340(6139), 1438-1441.

Cauchy, A.-L. (1821). *Cours d'analyse de l'École Royale Polytechnique.* Paris: Imprimerie Royale.

Demetrius, L. (2003). Quantum statistics and allometric scaling of organisms. *Physica A*, 322, 477-490.

Demetrius, L. (2006). The origin of allometric scaling laws in biology. *Journal of Theoretical Biology*, 243(4), 455-467.

Demetrius, L. (2010). Quantum metabolism and allometric scaling relations in biology. *Proceedings of the Royal Society A*, 466(2124), 3543-3561.

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

Eastwood, M. D. (2026). Paper I: The ARC Principle - Understanding as a Function of Intelligence and Recursive Depth. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper II: Experimental Validation of the ARC Principle Across Frontier Language Models. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper III: The Alignment Scaling Problem - Why External AI Safety Approaches Cannot Scale With Recursive Capability. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper IV.a: Baked-In vs Computed Alignment - A Three-Tier Empirical Hierarchy. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper IV.b: Alignment Saturation at Low Depth. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper IV.c: ARC-Align Benchmark - A Four-Layer Blinding Protocol for AI Alignment Evaluation. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper IV.d: The Effect of Blinding on AI Alignment Evaluation. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper V: The Stewardship Gene - A Developmental Alignment Architecture for Self-Modifying AI. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper VI: The Honey Architecture - Why Embedded Safety Prevents Collapse Under Recursive Self-Modification. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper VII: Cauchy Unification - ARC/Cauchy Scaling Classification Across 25 Domains. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). Paper VIII: The Load-Bearing Proof - Three Independent Experiments Testing Structural Entanglement Under the Eden Protocol. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). On the Origin of Scaling Laws: Cauchy Functional Equations as the Mathematical Foundation of Recursive Scaling. ARC/Eden Research Programme. OSF: 10.17605/OSF.IO/6C5XB.

Eastwood, M. D. (2026). The Eden Protocol: Engineering Specification for Embedded AI Alignment. ARC/Eden Research Programme.

Eastwood, M. D. (2026). Eden Protocol: Philosophical Vision. ARC/Eden Research Programme.

Glazier, D. S. (2008). Effects of metabolic level on the body size scaling of metabolic rate in birds and mammals. *Proceedings of the Royal Society B*, 275(1641), 1405-1410.

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

He, J. H. & Chen, W. X. (2003). Fractal estimation of cell biological systems. *Fractals*, 11, 437.

He, J. H. & Zhang, L. N. (2004). Fifth dimension of life and the 4/5 allometric scaling law for human brain. *Cell Biology International*, 28, 809-815.

Hu, E. J. et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022.* arXiv:2106.09685.

Maino, J. L., Kearney, M. R., Nisbet, R. M. & Kooijman, S. A. L. M. (2014). Reconciling theories for metabolic scaling. *Journal of Animal Ecology*, 83, 20-29.

Ouyang, L. et al. (2022). Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155.

Qwen Team (2024). Qwen 2.5 Technical Report. arXiv:2412.15115.

Sandberg, A. & Bostrom, N. (2008). Whole Brain Emulation: A Roadmap. *Future of Humanity Institute, Oxford University.* Technical Report 2008-3.

Weibel, E. R., Bacigalupe, L. D., Schmitt, B. & Hoppeler, H. (2004). Allometric scaling of maximal metabolic rate in mammals: muscle aerobic capacity as determinant factor. *Respiratory Physiology & Neurobiology*, 140(2), 115-132.

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.

Zhang, X. et al. (2025). Darwin Gödel Machine: Open-Ended Self-Improving AI. arXiv:2505.22954.

Zhao, J. (2022). Universal growth scaling law determined by dimensionality. *arXiv:2206.08094.*

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 |