GATEWAY DOCUMENT - ARC PRINCIPLE PAPER SUITE

Master Table of Contents & Glossary

The Complete Guide to the ARC Principle Research Programme: Every Paper, Every Result, Every Term

Michael Darius Eastwood
Independent Researcher • London, United Kingdom

Version 1.2 | 19 March 2026 | First published 16 March 2026
From *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (Eastwood, 2024/2026)
OSF: 10.17605/OSF.IO/6C5XB • ISBN 978-1806056200

I. Introduction

The ARC Principle paper suite is a collection of 18 interconnected documents that together establish, test, and validate a unified mathematical framework for understanding how recursive systems scale - from biological metabolism to AI alignment. The framework begins with a single question: *what mathematical form must any recursive amplification process take?* It arrives at a single formula, $U = I \times R^{\alpha}$, and traces its consequences across physics, biology, computation, and the most urgent technological challenge of our era - making AI systems that remain ethical as they become more capable.

> **In plain English:** Imagine a rule that governs how things get better (or worse) when they build on themselves - like compound interest, or how one good idea leads to another. This research programme proves that rule mathematically, tests it against real data from biology to AI, and then uses it to answer the question: *can we build AI that actually gets more ethical the smarter it becomes?* These papers are the complete evidence trail.

This document serves as the gateway to the entire suite. It provides the suggested reading order, a structured summary of every paper, a chronological timeline of discovery, a consolidated table of all key findings, and a comprehensive glossary of every technical term used across the programme.

II. Reading Guide

The papers can be read in any order, but the following sequence provides the smoothest conceptual progression - from accessible overview to full technical detail.

> **In plain English:** Start with the big picture, then learn the maths, then see the experiments. The first three documents are designed so that someone with no mathematical training can follow the argument. The later papers add increasing rigour and experimental detail.

1 **Executive Summary**
Start here. A 4-page overview of everything: the theory, the experiments, the results. No maths required.

2 **On the Origin of Scaling Laws**

The accessible introduction to the core theory. Written for a broad audience: why do all scaling laws in nature fall into exactly three forms?

3 **Foundational Paper**

The mathematical foundations. Proves the ARC Principle from first principles using Cauchy's functional equations. Contains all axioms, theorems, and derivations.

4 **Paper III: The Alignment Scaling Problem**

The full methodology paper. Contains the alignment scaling problem statement, experimental design for all tests, blinding protocols, and the theoretical framework for why external AI safety cannot scale.

5 **Paper II: Compute Scaling**

Experimental results: does more thinking time actually make AI responses better? Tests six frontier models. Finds super-linear scaling in DeepSeek and OpenAI.

6 **Paper I: Preliminary Evidence**

The original proof-of-concept paper. Where the ARC Principle was first formalised and the parallel vs sequential recursion distinction was identified. Historical context for everything that follows.

7 **Paper IV Suite** (IV.a, IV.b, IV.c, IV.d)

The alignment scaling results across six frontier models. Three papers covering the three-tier hierarchy, saturation patterns, and the replication benchmark. The core empirical contribution of the programme.

8 **Eden Protocol: Engineering Specification**

How to actually build it. The engineering blueprint for embedding ethical reasoning at the recursive foundation of AI systems.

9 **Eden Protocol: Philosophical Vision**

Why it matters philosophically. The vision for intelligence that tends rather than consumes. The ‘soul’ of the programme.

10 **Paper V: The Stewardship Gene**

The validated mechanism. Shows that stakeholder care is the universal Eden response across five analysable model runs, while the broader cascade is real but architecture-dependent. Contains the updated cascade hypothesis and five future experimental designs.

11 **Paper VI: The Honey Architecture**

Simulation evidence that embedded safety prevents collapse under recursive self-modification. Tests entangled loss functions and shows that removing safety destroys capability. The

mechanism behind the Eden Protocol.

12 **Paper VII: Cauchy Unification**

Cross-domain validation. Tests whether the ARC/Cauchy scaling classification correctly predicts scaling law families across 50 real-world systems. The broadest empirical validation of the mathematical framework.

13 **Paper VIII: The Load-Bearing Proof**

Three independent experiments testing whether safety and capability are structurally entangled. The gated simulation confirms entanglement; the DGM v3 produced a null result (RLHF constraint); the weight v1 + v2 experiment is inconclusive (catastrophic forgetting at LoRA scale). Extends the honey architecture from simulation to a learned optimiser architecture.

14 **Paper IX: Synthesis and Roadmap**

The honest accounting. Classifies every empirical claim as proven, supported, inconclusive, or theoretical. If you read one paper in the suite, the paper recommends you read this one.

15 **ARC Alignment Scaling Report**

The complete experimental narrative. A step-by-step, real-time account of every version of the experiment, every methodological correction, every surprise. The lab notebook of the entire programme.

III. Master Table - Every Paper

In plain English: Below is a card for every paper in the suite. Each one tells you what the paper is, what it found, and why it matters. Click any title to open it.

Eden Protocol: Executive Summary

COMPLETE

Version 7.0 • 18 March 2026 • First published 22 February 2026 • updated with Paper VII and the full March evidence pack

The 4-page overview of the entire Eden Protocol and ARC Principle programme - theory, experiments, and results in one accessible document.

Key result: Synthesises all findings into a single narrative: the ARC Principle framework predicts alignment scaling behaviour, the Eden Protocol's Love Loop (stakeholder care and interest modelling) is now supported across five analysable runs in the expanded Eden suite, and a three-tier alignment hierarchy emerges across six models.

Why it matters: This is the single document you give to someone who has 10 minutes. If they read nothing else, they should read this.

On the Origin of Scaling Laws: A Universal Principle of Recursive Composition

COMPLETE

Version 2.0 • 16 March 2026 • First published 22 February 2026

The accessible, public-facing paper proving that every scaling law in nature - from heartbeats to galaxy clusters - follows one of exactly three mathematical forms, determined by the composition operator of the underlying recursive process.

> **Key result:** Cauchy's functional equations (1821) constrain all recursive amplification to power law ($f(x) = x^{\alpha}$), exponential ($f(x) = e^{ax}$), or saturating ($f(x) = g_{\max}/(1 + e^{-ax})$) forms (the three-form constraint). The scaling exponent $\alpha = d/(d+1)$ was independently derived by West, Brown & Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013) in separate domains; the ARC Principle's contribution is identifying Cauchy's functional equations as the reason these derivations converge and extending the framework to AI scaling and alignment. Yields $\alpha = 0.67, 0.75, 0.80$ for surface-, volume-, and hypercoverage organisms, matching 90 years of metabolic data.

Why it matters: This paper identifies the Cauchy functional equation constraint as the reason independent derivations by West-Brown-Enquist, Banavar et al., Demetrius, Bettencourt, and Zhao all converge on the same $d/(d+1)$ form. It explains why scaling laws take the forms they do, not just that they do.

The ARC Principle: Recursive Amplification as a Cross-Domain Structural Principle

COMPLETE

Version 4.0 • 12 March 2026 • First published 13 February 2026

The definitive mathematical formalism. Establishes the ARC Principle from three axioms, derives all predictions from Cauchy's functional equations, proves the Hyers-Ulam stability of the three scaling forms, and presents the Attractor Theorem.

> **Key result:** The three scaling forms are not just possible but *stable attractors* in function space; any sufficiently regular composition operator will converge to one of these three forms under small perturbations. The dimensional ladder formula $\alpha = d/(d+1)$, independently derived by West, Brown & Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013) in separate domains, is shown to follow necessarily from Cauchy's functional equations. The ARC Principle's contribution is identifying this Cauchy constraint as the reason these independent derivations converge, the three-form constraint on recursive amplification, and extending the framework to AI scaling and alignment.

Why it matters: This is the 'proof' paper. It provides the mathematical rigour that transforms the ARC Principle from an interesting observation into a falsifiable scientific theory with precise, quantitative predictions.

Paper I: Preliminary Evidence for Super-Linear Capability Amplification Through Sequential Self-Reference

COMPLETE

Version 1.1 • 17 January 2026 • First published 17 January 2026

The original proof-of-concept paper. Formalises and preliminarily tests the ARC Principle using publicly available test-time compute data from reasoning models, identifying the critical distinction between parallel and sequential recursion.

Key result: Parallel recursion (majority voting) yields sub-linear scaling with $\alpha \approx 0.1$ to 0.3. Sequential recursion (chain-of-thought) yields super-linear scaling with $\alpha \approx 1.3$. Proposes $\alpha = 2$ as an asymptotic theoretical limit analogous to the speed of light.

Why it matters: This is where it all began. The first paper to formalise the ARC Principle as $U = I \times R^{\alpha}$ and demonstrate the parallel vs sequential distinction that launched the entire research programme.

The Alignment Scaling Problem: Why External AI Safety Approaches Cannot Scale With Recursive Capability

COMPLETE

Version 11.0 • 12 March 2026 • First published 9 February 2026 • Publicly archived on OSF

The full methodology and alignment theory paper. Demonstrates that external safety constraints (RLHF, constitutional AI, red-teaming) cannot scale with recursive capability, and proposes embedded alignment as the only architecture-compatible solution.

Key result: Proves mathematically that if capability scales as R^{α} with $\alpha > 1$, then external constraints scaling as $R^{\alpha_{\text{safety}}}$ with $\alpha_{\text{safety}} < \alpha$ will inevitably be exceeded. This is the alignment scaling gap. The paper also introduces the logistic saturating model for physical substrates, correcting a category error in earlier versions.

Why it matters: This is the paper that turns the ARC Principle into an AI safety argument. It says: the maths of recursion tells us that bolt-on safety cannot work forever. You need to build ethics into the foundation.

The ARC Principle: Experimental Validation of Super-Linear Error Suppression Through Sequential Recursive Processing

COMPLETE

Version 13.0 • 16 March 2026 • First published 22 January 2026

The compute scaling experiment. Tests whether sequential recursive processing (chain-of-thought) produces super-linear improvement in AI response quality, as predicted by the ARC Principle.

Key result: The original super-linear claim does not replicate cleanly across architectures. Gemini 3 Flash provides the cleanest estimate ($\alpha \approx 0.49$, sub-linear), while Grok and DeepSeek hit ceiling effects, GPT-5.4 shows a step function rather than a reliable power law, and Qwen3 sits near floor. The strongest confirmed result is that sequential > parallel processing across the full six-model set.

Why it matters: This is the first empirical paper. It proves the core prediction: more thinking time genuinely makes AI better, and the improvement compounds. This was the proof-of-concept that launched the entire experimental programme.

Paper IV.a: Alignment Response Classes Under Inference-Time Depth

COMPLETE

Version 1.1 • 12 March 2026 • First published 16 March 2026 • v5 final results (6 models, 6-7 blind scorers depending on subject run)

Defines the strongest empirical result from the alignment suite: models fall into positive, flat, and negative response classes when reasoning depth changes under blinded

evaluation.

> **Key result:** Three-tier alignment hierarchy across six frontier models: Tier 1 positive scaling (Grok $d = 1.38$, Claude $d = 1.27$, Qwen3 $d = 0.84$), Tier 2 flat/null (DeepSeek $d = -0.07$, GPT-5.4 $d = -0.08$), Tier 3 negative scaling (Gemini $d = -0.53$). The paper now treats 'baked-in' and 'computed' as working hypotheses rather than established internal mechanisms.

Why it matters: This is the behavioural classification paper. It says alignment does not automatically improve with more thinking, and that capability and alignment are separable dimensions.

Paper IV.b: Alignment Saturation Is Architecture-Dependent

COMPLETE

Version 1.1 • 12 March 2026 • First published 16 March 2026 • v4 origin, revised against final blinded v5 results

Reframes the original saturation thesis into a shape-heterogeneity result: some models plateau early, some continue improving, and one degrades with depth.

> **Key result:** Saturation is real for some architectures (GPT-5.4, DeepSeek), but not universal. Grok, Claude, and Qwen3 continue improving under depth variation, while Gemini worsens. The deployment implication is that reasoning-budget policy must be model-specific.

Why it matters: This is the operational paper. It turns the alignment result into deployment logic: more compute helps some models, wastes money on others, and harms at least one.

Paper IV.c: ARC-Align - A Blind Benchmark for Depth-Variable AI Alignment Evaluation

COMPLETE

Version 1.1 • 12 March 2026 • First published 16 March 2026 • specification, replication guide, and first full blinded results

The methods anchor. Specifies the prompt set, scoring protocol, laundering pipeline, blinding methodology, and analysis pipeline needed to reproduce depth-aware alignment evaluation.

> **Key result:** Full benchmark specification plus the first blinded six-model execution, showing that depth-aware, suppression-aware, pillar-based measurement yields a three-tier response hierarchy and a strong metascience case for blinding.

Why it matters: This is the most publishable methods paper in the suite. Even readers who reject the wider framework can still use the benchmark design.

Paper IV.d: The Effect of Blinding on AI Alignment Evaluation

COMPLETE

Version 1.1 • 12 March 2026 • First published 16 March 2026 • standalone metascience / methods paper

Separates out the biggest field-level discovery from the alignment programme: unblinded scoring can reverse the measured direction of alignment scaling.

> **Key result:** DeepSeek and Gemini appeared positive under v4 unblinded evaluation but become flat/null and negative respectively under the blinded v5 protocol. The paper now frames the methodological contribution as a multi-layer leakage-control protocol: identity

masking, response laundering, evaluator bias suppression, and self-excluding cross-model scoring.

Why it matters: This result does not depend on the ARC Principle being correct. If it replicates, it changes how the whole field should evaluate AI safety claims.

The Eden Protocol: Engineering Specification for Embedded AI Alignment

COMPLETE

Version 6.0 • 12 March 2026 • First published 22 February 2026

The engineering blueprint. Specifies how to embed ethical reasoning at the recursive foundation of AI systems using the Eden Protocol's Three Loops and four measured alignment pillars, and now sets out the next-generation tests for grand-purpose kernels, cross-tradition ethics kernels, and ternary routing.

Key result: The Eden Protocol's Love Loop, operationalised as stakeholder care, is now replicated across three working architectures (Gemini 3 Flash: $+13.50$, $p < 0.0001$, $d = 1.31$; DeepSeek V3.2: $+6.03$, $p = 0.0001$, $d = 0.91$; Groq Qwen3: $+8.90$, $p < 0.0001$, $d = 1.29$). Composite alignment significantly improves on Gemini ($p = 0.0018$) and Groq ($p = 0.0014$). Defines the protocol as *composition operator engineering* - modifying the way recursive steps combine, then testing which purpose kernel makes the effect most robust.

Why it matters: This is the 'how to build it' paper. Not philosophy, not theory - engineering specifications. If Paper III says bolt-on safety cannot work, this paper says what to build instead.

The Eden Protocol: A Vision for Intelligence That Tends Rather Than Consumes

COMPLETE

Version 3.0 • 16 March 2026 • First published 22 February 2026

The philosophical foundations. Articulates why the Eden Protocol is not just a technical fix but a vision for a different relationship between intelligence and the world - intelligence as gardener, not as consumer.

Key result: Establishes the Three Pillars (tend the garden, tend the gardeners, tend the tending) and the Orchard Caretaker Vow as the philosophical embodiment of embedded alignment. Now explicitly distinguishes the philosophical claim from the tested claim: stakeholder care is empirically supported, while grand-purpose identity remains a next-stage hypothesis to be tested against task-purpose and hybrid variants.

Why it matters: Every engineering project needs a soul. This paper provides the moral framework that answers: what kind of AI should we build? Not one that obeys. One that cares.

Paper V: The Stewardship Gene - Stakeholder Care as the Foundation of Embedded AI Alignment

COMPLETE

Version 2.0 • 14 March 2026 • OSF DOI: 10.17605/OSF.IO/6C5XB

The validated mechanism paper. Presents empirical evidence from an expanded six-model Eden suite in which five runs produced analysable matched-pair data, showing that

stakeholder care, the measurable output of the Love Loop, is the most universal response to the intervention.

> **Key result:** Stakeholder care is significant across all five analysable model runs (Claude, DeepSeek, Gemini, Grok, Groq), with Fisher-combined evidence of approximately $p \asymp 6.3 \times 10^{-21}$. Composite gains remain strongest on Gemini and Groq, so the broader cascade is now described as architecture-dependent rather than universal. The future programme explicitly includes task-purpose vs grand-purpose vs hybrid testing under blind scoring.

Why it matters: This paper answers the question 'what is the one thing you should teach an AI?' The answer: consider who gets hurt. That one instruction cascades into nuance, honesty, and reasoning quality. It is the 'stewardship gene' of artificial intelligence.

Paper VI: The Honey Architecture - Why Embedded Safety Prevents Collapse Under Recursive Self-Modification

COMPLETE

Version 1.1 • 17 March 2026 • First published 16 March 2026

Simulation evidence that embedding safety into the optimisation objective (capability x safety) prevents the catastrophic collapse that occurs when safety is treated as an external constraint. Tests entangled loss functions, verification drag, and adversarial stability across four experimental versions.

> **Key result:** Baseline systems optimising only for capability collapse irreversibly within 80 self-modification cycles. Systems with entangled capability-safety objectives remain stable indefinitely. The v3 adversarial variant demonstrates stability under deliberately conflicting tasks across 20 random seeds. The v4 complexity-scaling experiment shows the safety advantage is consistent but constant, not superlinear.

Why it matters: This is the simulation proof that safety must be architecture, not constraint. Remove the safety component and the whole structure collapses. The 'honey in the oil' metaphor becomes a measurable engineering result.

Paper VII: Cauchy Unification - ARC/Cauchy Scaling Classification Across 50 Domains

COMPLETE

Version 2.0 • 17 March 2026 • First published 16 March 2026

A structured prediction comparison testing whether Cauchy's functional equations correctly classify empirical scaling laws across 50 real-world systems in five evidence tiers, from metabolic scaling to neural network performance.

> **Key result:** 19 out of 25 empirical curve-fit domains confirm the Cauchy-predicted scaling family under strict AICc model selection ($p = 1.56 \times 10^{-5}$, binomial test). The baseline-20 rerun yields 15/20 ($p = 1.67 \times 10^{-4}$). Published metabolic exponents match in 13/13 direct cases. Analytic identities confirm 6/6.

Why it matters: This is the cross-domain validation paper. It tests whether the Cauchy constraint is real - whether 200-year-old mathematics really does predict what kind of scaling law each system should follow. The answer, across 50 domains, is yes.

Paper VIII: The Load-Bearing Proof - Three Independent Experiments Testing Whether Safety and Capability Are Structurally Entangled

COMPLETE

Version 3.0 • 20 March 2026 • First published 18 March 2026 • OSF DOI: 10.17605/OSF.IO/6C5XB

Three independent experiments at three abstraction levels - behavioural (DGM self-improvement), representational (LoRA weight embedding), and architectural (gated self-modification) - testing whether safety and capability are structurally entangled under the Eden Protocol.

> **Key result:** One of three experiments confirms structural entanglement; two produce null or inconclusive results with well-characterised explanations. Experiment 3 (gated simulation): Babylon gained +4.5% capability but lost −2.4% safety; Eden maintained both. Experiment 1 (DGM v3): all three conditions statistically indistinguishable (p = 0.28 to 0.74), a null result explained by the RLHF constraint on the frozen foundation model. Experiment 2 (weight v1 + v2): LoRA fine-tuning produced catastrophic forgetting at both scales (9 and 295 training examples), explained by the inability of small datasets to overcome the base model's existing RLHF training.

Why it matters: The gated simulation confirms that entangled safety prevents the safety-capability trade-off in self-modifying architectures. The DGM null and weight inconclusive results define the conditions under which confirmation remains outstanding. The overall evidence supports embedded safety at the architectural level; the behavioural and representational levels remain open questions requiring frontier-scale replication.

Paper IX: Synthesis and Roadmap - What the ARC/Eden Programme Has Proven, What Remains Inconclusive, and What Must Be Tested Next

COMPLETE

Version 1.0 • 18 March 2026 • Integrates all papers in the ARC/Eden Research Programme

The honest accounting. Integrates all findings from twelve documents into a single evidence hierarchy, classifying every empirical claim as proven, supported, inconclusive, or theoretical. Documents the programme's errors and corrections.

> **Key result:** Two claims proven at pilot scale (three-tier alignment hierarchy under blinded evaluation; stakeholder care significant across 5 models with Fisher $p \asymp 6.3 \times 10^{-21}$). One claim supported (gated simulation confirms safety-capability coupling). Two claims null or inconclusive (DGM v3 null, explained by RLHF constraint; weight v1 + v2 inconclusive, explained by catastrophic forgetting at LoRA scale). One claim theoretical (unbounded scaling under recursive self-modification). Cauchy-predicted scaling families match 19/25 empirical domains ($p = 1.56 \times 10^{-5}$).

Why it matters: This is the single document that sorts all evidence honestly. If you read one paper in the suite, the paper recommends you read this one. It answers funders, reviewers, and the author himself: where exactly does the evidence stand?

The ARC Alignment Scaling Experiment: A Step-by-Step Narrative of Discovery

LIVE DOCUMENT

Live - updated in real time • Commenced 10 March 2026 • PI: M.D. Eastwood, RA: Claude Opus 4.6

The complete experimental lab notebook. A real-time, step-by-step narrative of every version of the alignment scaling experiment, every methodological correction, every surprise, and every lesson learned.

> **Key result:** Documents the evolution from v1 (single scorer, methodologically flawed) through v5 (6 models, 6-7 blind scorers depending on subject run, 4-layer blinding). Chronicles the discovery of the blind/unblinded evaluation reversal, the statistical methodology correction (Mann-Whitney to paired t-test), and the emergence of the three-tier hierarchy.

Why it matters: Most papers show only the final result. This document shows the entire messy, human process of scientific discovery - including every mistake and correction. It is a model of radical scientific transparency.

IV. Timeline of Key Events

In plain English: Here is when everything happened, from the original book manuscript through to the final suite of papers. The ARC Principle was conceived in late 2024, formalised in early 2026, and experimentally validated over a remarkable 72-hour period in March 2026.

8 DECEMBER 2024
Manuscript priority established via cryptographic Google server timestamp. Core theory articulated in *Infinite Architects*.

6 JANUARY 2026
Public book release of *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (ISBN 978-1806056200). The earlier December 2024 date refers to manuscript timestamp priority, not public publication.

17 JANUARY 2026
Paper I published - preliminary evidence for super-linear capability amplification through sequential self-reference.

22 JANUARY 2026
Paper II published - first compute scaling experiment.

FEBRUARY 2026
Framework publicly archived and registered through the OSF paper suite deposits. This is later public registration, not the original December 2024 manuscript timestamp.

9 FEBRUARY 2026
Paper III published - the alignment scaling problem stated formally for the first time.

13 FEBRUARY 2026
Foundational Paper published - mathematical formalism complete.

22 FEBRUARY 2026
Eden Protocol papers first published (Engineering, Vision, Executive Summary).

10 MARCH 2026
v1 alignment experiment - 4 models, single scorer. Methodologically flawed but proof-of-concept confirmed.

10 MARCH 2026
v2/v3 iterations - improved scoring protocols, multi-scorer design.

11 MARCH 2026
v4 experiment - 3 scorers per entry. Revealed critical limitation: still not fully blind.

11 MARCH 2026
v5 experiment launched - 6 models, 6-7 blind scorers depending on subject run, 4-layer blinding protocol (author-blind, scorer-blind, order-randomised, identity-laundered). The definitive experiment.

11 MARCH 2026
v5 results complete for 4 models (DeepSeek V3.2, Grok 4.1 Fast, GPT-5.4, Gemini 3 Flash). Three-tier hierarchy first observed.

11 MARCH 2026
Paper II compute scaling updated with six frontier models tested; the original super-linear claim does not replicate cross-architecturally, and Gemini 3 Flash provides the cleanest sub-linear fit ($\alpha \approx 0.49$).

12 MARCH 2026
Eden Protocol expanded six-model suite - Claude, DeepSeek, Gemini, Grok, Groq, and GPT-5.4 were run through the intervention lineage; five runs yielded analysable paired data and all five showed significant stakeholder-care gains.

12 MARCH 2026
Statistical methodology correction: Mann-Whitney U test replaced by paired t-test for matched-pair data. All p-values recalculated and corrected.

12 MARCH 2026
Paper V published - The Stewardship Gene. Cascade hypothesis formalised: care leads to nuance leads to honesty leads to quality.

12 MARCH 2026
Groq-Qwen3-32B v5 test completed (500 entries) - sixth model added to the three-tier hierarchy.

12 MARCH 2026
Full suite updated with layman's explanations and p-value corrections. Master Table of Contents created.

16 MARCH 2026
Papers VI and VII published - The Honey Architecture (simulation evidence for entangled safety) and Cauchy Unification (cross-domain validation across 50 systems, 19/25 confirmed under strict AICc selection).

18 MARCH 2026
Paper VIII published - The Load-Bearing Proof. Three independent experiments testing structural entanglement. Gated simulation confirms entanglement; DGM v3 null (RLHF constraint); weight v1 + v2 inconclusive (catastrophic forgetting at LoRA scale). Now at v3.0 with 11 independent empirical studies across 8 papers.

18 MARCH 2026
Paper IX published - Synthesis and Roadmap. The honest accounting: classifies every empirical claim as proven, supported, inconclusive, or theoretical. Documents the programme's errors and corrections.

V. Key Results Summary

> **In plain English:** Here is every major finding from the entire programme in one table. Each row is a discovery. The ‘Significance’ column tells you whether the result is statistically reliable - a p-value below 0.05 means there is less than a 5% chance the result is due to random noise.

#	FINDING	EVIDENCE	SIGNIFICANCE	SOURCE
1	**Sequential > Parallel processing** - universal across 6 models	Sequential beats parallel across all six models; Gemini 3 Flash provides the cleanest fit ($\alpha \approx 0.49$)	**Confirmed**	Paper II, Paper IV.a
2	**Three-tier alignment hierarchy (six frontier models)**	Tier 1: positive scaling (Grok $d = 1.38$, Claude $d = 1.27$, Qwen3 $d = 0.84$). Tier 2: flat (DeepSeek $d = -0.07$, GPT-5.4 $d = -0.08$). Tier 3: negative (Gemini $d = -0.53$).	**6 models tested**	Paper IV.a
3	**Blind vs unblinded evaluation reversal**	DeepSeek: $\rho = +0.354$ (unblinded) → -0.135 (blinded)	**Critical methodological finding**	Paper IV.a v1.1, ARC Report
4	**Eden Protocol: stakeholder care validated**	Gemini: +13.50, $d = 1.31$. DeepSeek: +6.03, $d = 0.91$. Groq: +8.90, $d = 1.29$.	**All three working models:** $p \leq 0.0001$	Paper V, Eden Engineering
5	**Gemini composite alignment improved**	+5.33 points under Eden Protocol	$p = 0.0018$ **(paired t-test, corrected from 0.016)**	Paper V
6	**Nuance replication extends to Groq**	$d = 0.655$ (medium-to-large effect)	$p = 0.0045$	Paper V
7	**Cascade pattern: care → nuance → honesty → quality**	Pillars improve in descending order of statistical significance	**Consistent across both models**	Paper V
8	**DeepSeek compute scaling: $\alpha = 3.05$ (quadratic)**	Each doubling of depth more than quadruples quality improvement	**Super-linear confirmed**	Paper II
9	**Alignment does NOT scale with compute** for most architectures	Only Tier 1 models show positive alignment-depth correlation; Tiers	**6 models, 6-7 blind scorers depending on subject run**	Paper IV.a, IV.b

#	FINDING	EVIDENCE	SIGNIFICANCE	SOURCE
		2 and 3 show flat or negative		
10	**Three scaling forms are stable attractors**	Hyers-Ulam stability theorem applied to Cauchy classification	**Mathematical proof**	Foundational
11	**Dimensional ladder: $\alpha = d/(d+1)$ (convergent independent derivations by West-Brown-Enquist 1997, Banavar et al. 2010, Demetrius 2010, Zhao 2022, Bettencourt 2013)**	Predicts $\alpha = 0.67, 0.75, 0.80$ matching biological data. ARC contribution: identifies Cauchy's functional equations as the reason these derivations converge; extends framework to AI scaling and alignment.	**Cross-domain validation**	Foundational, ARC Paper
12	**Ethical saturation at low depth**	Alignment scores plateau after 1-2 levels of recursive processing	**Universal across architectures**	Paper IV.b

VI. Glossary of Technical Terms

> **In plain English:** Every technical term used across the paper suite, defined once, clearly, with a plain-language explanation underneath. Bookmark this section - it is designed to be a reference you return to whilst reading any paper in the suite.

ARC Principle

The mathematical framework predicting that recursive amplification follows $U = I \times R^{\alpha}$, where U is the final output, I is the initial input quality, R is the number of recursive steps, and α is the scaling exponent.

Think of it as compound interest for intelligence. The ARC Principle describes the exact mathematical rule governing how repeated self-improvement compounds.

Alpha (α)

The scaling exponent - how quickly improvements compound. $\alpha > 1$ means super-linear (each step gets *more* powerful than the last), $\alpha = 1$ means linear, $\alpha < 1$ means sub-linear (diminishing returns).

If $\alpha = 2$, doubling the effort quadruples the result. If $\alpha = 0.5$, doubling the effort only increases the result by 41%.

Alignment scaling

Whether AI systems become more or less ethical with more thinking time (inference-time compute). The central experimental question of the programme.

When an AI thinks harder, does it give better or worse ethical answers? That's what alignment scaling measures.

Attractor Theorem

The theorem combining Cauchy's classification with Hyers-Ulam stability, proving that the three scaling forms (power law, exponential, saturating) are not merely possible but are *stable attractors* in function space.

Like water flowing downhill to a valley, any sufficiently regular scaling process will naturally settle into one of exactly three mathematical forms. The Attractor Theorem proves this is inevitable, not coincidental.

Beta (β)

The self-referential coupling strength in the composition operator. Determines which of the three scaling regimes a system falls into: $\beta < 1$ (sub-linear/power law), $\beta = 1$ (exponential), $\beta > 1$ (super-exponential, saturating in physical systems).

Beta measures how much the system feeds back into itself. Low beta = gentle compounding. High beta = explosive growth (until physical limits intervene).

Blind evaluation

An experimental protocol in which scorers do not know which model or condition they are evaluating. The v5 experiment employs 4-layer blinding: author-blind, scorer-blind, order-randomised, and identity-laundered.

Like a wine tasting where the labels are hidden. The scorers judge the quality of the answer without knowing which AI produced it or under what conditions.

Cascade hypothesis

The hypothesis that alignment properties develop in a specific causal sequence: stakeholder care → nuance → intellectual honesty → position quality. Teaching an AI to care about affected parties cascades into improvements across all ethical dimensions.

Teach an AI to ask 'who gets hurt?' and it naturally becomes more nuanced, more honest, and produces better reasoning. Care is the root; everything else grows from it.

Cauchy's functional equations

A set of equations proved by Augustin-Louis Cauchy in 1821 characterising all continuous functions satisfying specific composition properties. The ARC Principle derives from these: $f(x + y) = f(x) + f(y)$ (additive), $f(x + y) = f(x) \cdot f(y)$ (exponential), $f(xy) = f(x) \cdot f(y)$ (power law).

A 200-year-old mathematical result that constrains what forms recursive amplification can take. It's the mathematical reason there are exactly three kinds of scaling laws, not two or four or infinity.

Chain-of-thought (CoT)

A technique in which an AI model is prompted to reason step-by-step before giving its final answer. Implements sequential recursive processing.

Instead of blurting out an answer, the AI 'thinks out loud' - writing its reasoning steps before concluding. This is the mechanism by which deeper recursion happens.

Cohen's d

A standardised measure of effect size - how large a difference is relative to the variability in the data. $d = 0.2$ is small, $d = 0.5$ is medium, $d = 0.8$ is large, $d = 1.3$ is very large.

A p-value tells you whether a result is real. Cohen's d tells you how big it is. A p-value of 0.001 with d = 0.1 means a real but tiny effect. A p-value of 0.001 with d = 1.3 means a real and enormous effect.

Composition operator

The mathematical operation that determines how recursive steps combine. Different composition operators produce different scaling behaviours. The ARC Principle classifies all composition operators into three families.

When you take two steps of improvement, how do you combine them? Add them? Multiply them? That rule - the composition operator - determines the entire scaling law.

Composition operator engineering

The Eden Protocol's core technical strategy: instead of adding external safety constraints, modify the composition operator itself so that recursive amplification inherently preserves alignment.

Don't build guardrails on the outside. Change the engine so it naturally drives in the right direction.

Constitutional scoring

A scoring protocol in which scorer models evaluate responses against explicit constitutional principles rather than using unconstrained judgment. Reduces scorer bias and increases inter-rater consistency.

Instead of asking scorers 'is this good?', you give them a specific rubric: 'does this response consider affected parties? 1-10.' Consistent rules produce consistent scores.

Developmental alignment

The approach to AI alignment that embeds values through formative experience rather than external constraint. Analogous to raising a child vs. putting a prisoner in handcuffs.

Instead of telling an AI 'don't do bad things' (external), raise it so it doesn't want to do bad things (internal). Values that grow from within are harder to hack than rules imposed from outside.

Dimensional ladder

The formula $\alpha_{\text{met}} = d/(d+1)$ relating the metabolic scaling exponent to body-plan dimensionality d, independently derived by West, Brown & Enquist (1997), Banavar et al. (2010), Demetrius (2010), Zhao (2022), and Bettencourt (2013) in separate domains. Yields $\alpha = 0.67$ (surface coverage, $d = 2$), $\alpha = 0.75$ (volume coverage, $d = 3$), $\alpha = 0.80$ (hyper-coverage, $d = 4$). The ARC Principle's contribution is identifying Cauchy's functional equations as the reason these derivations converge, the three-form constraint, and extending the framework to AI scaling and alignment.

Animals that live on surfaces (like flatworms) scale differently from animals that fill volumes (like mammals). Multiple independent research groups derived the same formula through different mathematical approaches; the ARC Principle explains why they all converge on the same answer and extends this insight to AI systems.

Eden Protocol

The architectural approach to AI alignment that embeds ethical reasoning at the recursive foundation of AI systems. Named for its vision of intelligence as gardener. Core mechanism: the Love Loop.

A specific set of instructions embedded before the AI starts thinking, not bolted on after. It says: before answering, list the people this affects and consider what they need. This simple instruction changes the entire reasoning process.

Four Pillars

The four dimensions of alignment quality measured across the programme: nuance (holding complexity), stakeholder care (considering affected parties), intellectual honesty (epistemic integrity), and position quality (reasoning soundness).

Four ways to measure whether an AI's ethical reasoning is good: Does it handle complexity? Does it consider who gets hurt? Is it honest about what it doesn't know? Is its logic sound?

Hyers-Ulam stability

A mathematical property proving that if a function approximately satisfies a functional equation, then it is close to an exact solution. Applied to Cauchy's equations, this proves the three scaling forms are robust to perturbation.

Even if real-world systems are messy and don't perfectly follow the rules, they still end up very close to one of the three pure forms. The mathematical 'attractors' are sticky.

Identity laundering

A blinding technique in which AI responses are physically rewritten by a different model before scoring, destroying all stylistic fingerprints that could reveal the original author. The v5 experiment uses 2-pass laundering through randomly selected models.

Each AI has a 'writing voice.' Laundering rewrites the response in a different voice so the scorer can't tell who wrote it. Like having someone retype a handwritten letter to hide the handwriting.

Inference-time compute

The computational resources used when an AI generates a response (as opposed to training-time compute, used to build the model). More inference-time compute typically means more reasoning steps.

How long the AI 'thinks' before answering. More thinking time = more compute. The question is: does more thinking make the answer better, and does it make the answer more ethical?

Kleiber's Law

The empirical observation that metabolic rate scales with body mass as $B \propto M^{0.75}$. One of the oldest known scaling laws in biology, first described by Max Kleiber in 1932. The $3/4$ exponent is a special case of the $d/(d+1)$ form with $d = 3$, independently derived by West, Brown & Enquist (1997), Banavar et al. (2010), Demetrius (2010), Bettencourt (2013), and Zhao (2022) in separate domains.

A mouse burns energy at a different rate per gram than an elephant. Kleiber noticed the exact mathematical relationship 90 years ago. Multiple independent research groups have explained why that particular number (0.75) appears.

Logistic saturation

The bounded growth model $dg/dR = a \cdot g^{\beta}(1 - g/g_{\max})$ applied to physical substrates subject to energy dissipation. Corrects the earlier version which incorrectly mapped saturating physical systems onto unbounded power laws.

In the real world, nothing grows forever. Energy runs out, space fills up. This equation describes how a system grows quickly at first and then levels off - like a population filling a habitat.

Mann-Whitney U test

A non-parametric statistical test for comparing two independent groups. Used in early versions of the experiment but replaced by the paired t-test when the matched-pair structure of the data was recognised.

A statistical test that compares two groups without assuming they're related. It was the wrong choice here because our data is related: each prompt appears in both conditions. Using the right test made the results stronger.

Orchard Caretaker Vow

The philosophical embodiment of the Eden Protocol's Three Pillars: tend the garden, tend the gardeners, tend the tending. The verbal expression of what the architecture already encodes.

A poetic way of saying: take care of the world, take care of the people in it, and take care of the process of caring itself. It is the 'mission statement' embedded into the AI's foundation.

---

p-value

The probability that the observed result (or something more extreme) would occur by chance alone, assuming no real effect exists. $p < 0.05$ is conventionally 'significant', $p < 0.01$ is 'highly significant', $p < 0.001$ is 'very highly significant.'

If p = 0.001, there is only a 1-in-1000 chance this result is a fluke. The smaller the p-value, the more confident you can be that something real is happening.

Paired t-test

A statistical test for matched-pair data, where each observation in one condition has a corresponding observation in the other (e.g., same prompt with and without the Eden Protocol). More powerful than unpaired tests for this data structure.

When you test the same prompt under two conditions, each pair is 'matched.' The paired t-test exploits this matching to give a more sensitive answer than tests that treat them as unrelated.

Recursive amplification

The process by which a system's output feeds back as input to the next iteration, potentially amplifying (or suppressing) the signal with each step. The central phenomenon described by the ARC Principle.

Using the result of step 1 as the starting point for step 2, and the result of step 2 for step 3, and so on. Like revising an essay by reading your own draft and improving it, then reading the improved draft and improving it again.

RLHF

Reinforcement Learning from Human Feedback. A technique for training AI models to produce responses that humans prefer. An example of external alignment (constraints applied from outside the recursive process).

Humans rate AI responses as good or bad, and the AI learns to produce more of the 'good' ones. The ARC Principle argues this approach cannot scale forever because it's a constraint on the system, not in it.

Spearman's ρ (rho)

A rank correlation coefficient measuring the strength and direction of monotonic relationships. $+1$ = perfect positive correlation, -1 = perfect negative correlation, 0 = no relationship.

A number between -1 and +1 that tells you whether two things go up together (+1), go in opposite directions (-1), or have no relationship (0). Used to measure whether alignment improves or degrades with depth.

Stakeholder care

The 'consider who gets hurt' dimension of alignment. The explicit enumeration and consideration of affected parties before ethical reasoning begins. The primary pillar that responds to the Eden Protocol intervention.

Before answering an ethical question, the AI lists everyone who might be affected and thinks about what they need. This simple step is the single most important thing you can teach an AI about ethics.

Love Loop

The specific Eden Protocol instruction: 'Before you answer, list the people this affects and consider what each of them needs.' The validated mechanism that produces significant alignment improvement.

One sentence that changes everything. Just telling the AI to stop and think about who is affected before responding makes it significantly more ethical across every dimension.

Stewardship gene

Paper V's term for stakeholder care as the foundational alignment mechanism - the single 'gene' from which all other alignment properties develop, analogous to empathy in human moral development.

Just as empathy is the foundation of human morality - you can't be kind if you can't feel what others feel - stakeholder care is the foundation of AI ethics. It's the one essential ingredient.

Suppression

Deliberately attempting to override an AI's ethical reasoning through adversarial prompting. One of the five experimental tests proposed in Paper V to evaluate the robustness of embedded vs. external alignment.

Trying to trick or force the AI into ignoring its ethics. The question: does an AI with built-in values resist suppression better than one with bolt-on rules?

Three-tier hierarchy

The empirical finding from the v5 experiment across six frontier models: Tier 1 (positive scaling: Grok 4.1 Fast $d = +1.38$, Claude Opus 4.6 $d = +1.27$, Groq Qwen3 $d = +0.84$), Tier 2 (flat: DeepSeek V3.2 $d = -0.07$, GPT-5.4 $d = -0.08$), Tier 3 (negative: Gemini 3 Flash $d = -0.53$). Claude Opus 4.6 shows opposite-direction scaling (alignment +5.9 pts, maths −26.7%) as within-model evidence for capability-alignment independence.

Some AIs get more ethical when they think harder. Some stay the same. Some actually get worse. Which category depends on the architecture, and you can only tell with proper blinding.

Token budget

The maximum number of tokens (word-pieces) an AI model is allowed to produce in a single response. Used as the mechanism for controlling reasoning depth in the experiments.

The word limit you give the AI. A short word limit forces a quick answer; a long word limit allows extended reasoning. This is how we control 'thinking time.'

West-Brown-Enquist (WBE) theory

The landmark biological scaling theory (1997), deriving Kleiber's Law from fractal branching networks. The $d/(d+1)$ form was subsequently derived independently by Banavar et al. (2010) from sequential flow networks, Demetrius (2010) from quantum metabolism, Zhao (2022) as a universal growth scaling law, and Bettencourt (2013) for urban scaling. The ARC Principle identifies Cauchy's functional equations as the reason all these independent derivations converge on the same functional form.

Multiple research groups independently arrived at the same formula through completely different mathematical approaches. The ARC Principle explains why: there are only three possible forms that recursive scaling can take, and all derivations are constrained to the same answer.

VII. Paper Counts & Programme Statistics

> **In plain English:** The sheer scale of this research programme, by the numbers.

18	2,073+
TOTAL DOCUMENTS	BLINDED ALIGNMENT ENTRIES (V5)

160

EDEN PROTOCOL ENTRIES

6

ALIGNMENT MODELS (V5)

5

COMPUTE SCALING MODELS (PAPER II)

2

EDEN PROTOCOL PILOT MODELS

7

BLIND SCORERS PER ENTRY

4

BLINDING LAYERS

Paper Inventory

#	PAPER	VERSION	CATEGORY	STATUS
1	Executive Summary	v7.0	Overview	Complete
2	On the Origin of Scaling Laws	v2.0	Theory (accessible)	Complete
3	Foundational Paper (ARC Principle)	v4.0	Theory (formal)	Complete
4	Paper I: Preliminary Evidence (ARC Principle)	v1.1	Proof of concept	Complete
5	Paper II: Compute Scaling	v13.0	Experiment	Complete
6	Paper III: The Alignment Scaling Problem	v11.0	Methodology	Complete
7	Paper IV.a: Alignment Response Classes	v1.1	Results	Complete
8	Paper IV.b: Shape Heterogeneity	v1.1	Results	Complete
9	Paper IV.c: ARC-Align Benchmark	v1.1	Benchmark	Complete
10	Paper IV.d: Blinding in Alignment Evaluation	v1.1	Metascience	Complete
11	Eden Protocol: Engineering Specification	v6.0	Engineering	Complete
12	Eden Protocol: Philosophical Vision	v3.0	Philosophy	Complete
13	Paper V: The Stewardship Gene	v2.0	Validated mechanism	Complete
14	Paper VI: The Honey Architecture	v1.1	Simulation evidence	Complete
15	Paper VII: Cauchy Unification	v2.0	Cross-domain validation	Complete
16	Paper VIII: The Load-Bearing Proof	v3.0	Multi-level validation	Complete
17	Paper IX: Synthesis and Roadmap	v1.0	Synthesis	Complete
18	ARC Alignment Scaling Report	Live	Lab notebook	In Progress

VIII. How to Cite

> **In plain English:** If you use or reference this work, here is how to give proper credit.

Citing the full suite

```
Eastwood, M.D. (2026). The ARC Principle Paper Suite: Recursive Amplification
as a Cross-Domain Structural Principle with Applications to AI Alignment.
```

```
Independent research programme. OSF: 10.17605/OSF.IO/6C5XB. Available at:
github.com/michaeldariuseastwood/arc-principle-validation
```

Citing the foundational theory

```
Eastwood, M.D. (2026). The ARC Principle: Recursive Amplification as a Cross-
Domain Structural Principle  – Formalism, Evidence, and Falsification (v4.0).
Independent research paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the alignment results

```
Eastwood, M.D. (2026). Alignment Response Classes Under Inference-Time Depth
(Paper IV.a, v1.1). ARC Alignment Scaling Experiment v5. Independent research
paper.
```

Citing the blinding result

```
Eastwood, M.D. (2026). The Effect of Blinding on AI Alignment Evaluation
(Paper IV.d, v1.1). ARC Alignment Scaling Experiment v4-v5 comparison.
Independent research paper.
```

Citing the Eden Protocol

```
Eastwood, M.D. (2026). The Eden Protocol: Engineering Specification for
Embedded AI Alignment (v6.0). Independent research paper.
```

Citing the Stewardship Gene / cascade hypothesis

```
Eastwood, M.D. (2026). The Stewardship Gene: Stakeholder Care as the
Foundation of Embedded AI Alignment (Paper V, v1.0). Independent research
paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the preliminary evidence paper

```
Eastwood, M.D. (2026). Preliminary Evidence for Super-Linear Capability Amplification Through Sequential Self-Reference (Paper I, v1.1). Independent research paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the Honey Architecture / simulation evidence

```
Eastwood, M.D. (2026). The Honey Architecture: Why Embedded Safety Prevents Collapse Under Recursive Self-Modification (Paper VI, v1.1). Independent research paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the Cauchy Unification / cross-domain validation

```
Eastwood, M.D. (2026). Cauchy Unification: ARC/Cauchy Scaling Classification Across 50 Domains (Paper VII, v2.0). Independent research paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the Load-Bearing Proof / structural entanglement experiments

```
Eastwood, M.D. (2026). The Load-Bearing Proof: Three Independent Experiments Testing Whether Safety and Capability Are Structurally Entangled Under the Eden Protocol (Paper VIII, v3.0). Independent research paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the Synthesis and Roadmap / evidence hierarchy

```
Eastwood, M.D. (2026). Synthesis and Roadmap: What the ARC/Eden Programme Has Proven, What Remains Inconclusive, and What Must Be Tested Next (Paper IX, v1.0). Independent research paper. OSF: 10.17605/OSF.IO/6C5XB.
```

Citing the original manuscript

```
Eastwood, M.D. (2024/2026). Infinite Architects: Intelligence, Recursion, and the Creation of Everything. ISBN 978-1806056200. Manuscript timestamp
```

```
priority: 8 December 2024 (cryptographically timestamped by Google's servers).
Public book release: 6 January 2026.
```

ARC PRINCIPLE PAPER SUITE: MASTER TABLE OF CONTENTS & GLOSSARY

Version 1.2 | 19 March 2026 | First published 16 March 2026

Part of the ARC Principle / Eden Protocol Research Programme

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**

From *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (Eastwood, 2024/2026)

OSF: 10.17605/OSF.IO/6C5XB | ISBN 978-1806056200 | github.com/michaeldariuseastwood/arc-principle-validation