RESEARCH PAPER

Paper II: Experimental Validation

Michael Darius Eastwood

First published 2026-01-22 · Updated 2026-03-13

ABSTRACT

Experimental paper showing the durable directional result that sequential recursion outperformed parallel sampling under the tested conditions, with later narrowing of the universal exponent claim.

RELATED READING

- Paper IV.d: The Effect of Blinding on AI Alignment Evaluation
- Paper IV.c: ARC-Align Benchmark
- Paper IV.a: Architecture-Dependent Alignment Response Classes

EASTWOOD'S ARC PRINCIPLE - PAPER II (EXPERIMENTAL VALIDATION)

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

A Mathematical Framework for Intelligence Amplification with Cross-Domain Convergent Evidence

Michael Darius Eastwood
Independent Researcher
Author, Infinite Architects: Intelligence, Recursion, and the Creation of Everything (2026)
London, United Kingdom | OSF: 10.17605/OSF.IO/8FJMA | ISBN 978-1806056200

Version 13.0 | March 2026 | Previous: v12.0 (11 March 2026), Definitive Final v11 (22 January 2026)
Manuscript Priority: 8 December 2024 (DKIM-verified) | Paper I Published: 17 January 2026
Correspondence: michael@michaeldariuseastwood.com | Repository: github.com/michaeldariuseastwood/arc-principle-validation

ABSTRACT

This paper presents experimental validation of the ARC Principle (Artificial Recursive Creation), a mathematical framework proposing that error rates in intelligent systems decrease according to a power law with recursive depth. The principle, first articulated in *Infinite Architects* (Eastwood, December 2024) and formalised in Paper I (Eastwood, 17 January 2026), predicts that the form of recursion determines the scaling regime: sequential recursion should yield super-linear error suppression (scaling exponent $\alpha > 1$), while parallel recursion should yield sub-linear suppression ($\alpha < 1$).

We conducted controlled experiments in two phases. Phase 1 (v11) used DeepSeek R1 with visible reasoning tokens on 12 competition-level mathematics problems, finding $\alpha_{\text{sequential}} \approx 2.24$ (95% CI: 1.5–3.0) and $\alpha_{\text{parallel}} \approx 0.0$. Phase 2 (v13) extended testing to six frontier models on 18 AIME/Putnam-level problems

(n=54 per depth per model) with bootstrap confidence intervals and 4-layer cross-verification.

Cross-architecture replication: The original $\alpha \approx 2.24$ (quadratic) does not replicate across architectures. Only Gemini 3 Flash produced clean, monotonic scaling data: $\alpha_{\text{seq}} = 0.49$ (regression, $r^2 = 0.86$, SE = 0.20, boot CI [−1.3, 2.9]). DeepSeek R1 reached ceiling (94.4%–100%). GPT-5.4 exhibited a binary step function (50% → 100%). Grok 4.1 Fast achieved 100% at all depths. Qwen3 showed no trend (~50%).

Parallel scaling confirmed universally: $\alpha_{\text{parallel}} \approx 0$ for every model tested, the strongest replicated finding. Sequential outperformed parallel for every model where both were measurable.

Revised parameter estimate: The most robust cross-architecture estimate is $\alpha_{\text{sequential}} \approx 0.49$ (sub-linear, from Gemini 3 Flash), substantially below the v11 single-model estimate of 2.24. Under the Intelligence Formula $\alpha = 1/(1-\beta)$ from Paper I, $\alpha < 1$ places current models in the physical regime (multiplicative composition through finite-dimensional networks) rather than the intelligence regime (recursive self-reference with $\alpha > 1$). The fundamental inequality $\alpha_{\text{sequential}} > \alpha_{\text{parallel}}$ remains confirmed.

> *New in v13.0* V13
>
> This version reports **complete tier-2 results** from the multi-model replication study:
>
> - **Six frontier AI models tested on 18 AIME/Putnam-level problems:** Grok 4.1 Fast, Claude Opus 4.6, Groq Qwen3, DeepSeek V3.2, GPT-5.4, and Gemini 3 Flash (n=54 per depth per model, 3 repeats)
> - **The quadratic claim ($\alpha \approx 2.24$) does not replicate across architectures:** Gemini 3 Flash provides the cleanest data with $\alpha_{\text{seq}} \approx 0.49$ (sub-linear, not quadratic)
> - **$\alpha_{\text{parallel}} \approx 0$ confirmed universally:** The strongest replicated finding across all models
> - **Token measurement bug identified and fixed:** `reasoning_tokens` → `total_tokens` corrected for GPT-5.4 and Qwen3
> - **Cross-verification complete:** All 3 disputed answers (ARC16=29, ARC17=176, ARC29=800) confirmed correct by hand
> - **Integration with alignment scaling:** Capability scaling (Paper II) and alignment scaling (Paper IV) shown to be independent dimensions
>
> All v11 findings are preserved in their original form. Conclusions have been revised to reflect the cross-architecture evidence.

Cross-domain evidence strengthens these findings. Google's Willow quantum chip (December 2024) demonstrated recursive error suppression with $\Lambda = 2.14$. Biological scaling laws show quarter-power exponents across 27 orders of magnitude via fractal recursive networks. The COGITATE consciousness study (*Nature*, April 2025) identified recurrent processing as the common denominator across theories.

The implications for AI safety depend critically on whether $\alpha > 1$ is achievable. If it is, alignment properties embedded in the reasoning process would scale super-linearly with capability while external constraints remain constant. Even with $\alpha < 1$, the directional finding that sequential reasoning improves capability supports the Eden Protocol from *Infinite Architects*: AI systems benefit from values embedded in reasoning rather than constraints imposed externally. However, the v13 data shows that capability and alignment are independent dimensions; more reasoning does not automatically mean better alignment.

Keywords: scaling laws, recursive intelligence, test-time compute, error suppression, AI safety, alignment, chain-of-thought reasoning, Eden Protocol, cross-domain validation, multi-model replication

1. INTRODUCTION

1.1 Background and Motivation

The scaling laws governing artificial intelligence have transformed our understanding of capability emergence. Kaplan et al. (2020) established power-law relationships between model performance and training compute, while Hoffmann et al. (2022) refined these with compute-optimal prescriptions. These foundational works revolutionised training methodology but address only pre-training scaling. They do not explain why allocating additional computation at inference time produces dramatic capability improvements, nor why different forms of such computation yield fundamentally different outcomes.

The emergence of reasoning models in late 2024 introduced test-time compute as a critical variable. OpenAI's o1 (September 2024) and DeepSeek's R1 (January 2025) allocate computational resources during inference to reason before responding. On mathematical reasoning benchmarks, these systems achieve performance previously thought to require order-of-magnitude larger models. Yet the mechanisms underlying this improvement remain incompletely characterised.

Two paradigms have emerged for allocating test-time compute:

Parallel recursion. Generate multiple independent solutions and select the best via majority voting or verifier scoring. This approach is computationally straightforward but, as documented by Brown et al. (2024), produces diminishing returns following sub-linear power laws.

Sequential recursion. Generate extended reasoning chains where each step builds explicitly on previous steps. Errors can be detected and corrected iteratively through self-reference. This approach produces compounding returns, but the scaling relationship has not been formally characterised; this paper addresses that gap.

1.2 The Research Question

Why does sequential reasoning dramatically outperform parallel sampling at equivalent computational cost? What mathematical principle governs this difference? And what are the implications for aligning increasingly capable AI systems?

1.3 Contribution of This Paper

This paper makes eight contributions:

1. **Mathematical formalisation.** We propose the ARC Principle: $E(R) = E_0 \times R^{-\alpha}$, where error rate E decreases from baseline E_0 as recursive depth R increases, governed by scaling exponent α. The form of recursion determines α.
2. **Controlled experimental validation.** Using DeepSeek R1 with visible reasoning tokens, we conduct the first compute-matched comparison between sequential and parallel recursion with direct measurement of recursive depth.
3. **Quantitative parameter estimation.** Single-model estimate: $\alpha \approx 2.2$ (v11, DeepSeek R1). Cross-architecture estimate: $\alpha \approx 0.49$ (v13, Gemini 3 Flash, $r^2 = 0.86$). Parallel: $\alpha \approx 0.0$ (universal). V13
4. **Converging evidence synthesis.** Combined with published data from OpenAI o1 and DeepSeek R1, multiple independent sources confirm $\alpha_{\text{seq}} > \alpha_{\text{par}}$.
5. **Cross-domain validation.** We demonstrate that recursive error suppression appears across quantum physics (Willow $\Lambda = 2.14$), biology (quarter-power scaling), and consciousness (COGITATE recurrence).
6. **AI safety implications.** We derive conditional alignment theorems, now qualified by the empirical finding that $\alpha < 1$ cross-architecturally.
7. **Cross-architecture replication.** Six frontier models tested on 18 AIME/Putnam-level problems with bootstrap CIs and 4-layer cross-verification. Four distinct scaling behaviours identified (ceiling, monotonic, step function, floor). V13
8. **Integration with alignment scaling.** Capability and alignment shown to be independent scaling dimensions across six frontier models, with a three-tier alignment hierarchy set by training methodology. V13

1.4 Priority Establishment

The ARC Principle was first articulated in the manuscript of *Infinite Architects: Intelligence, Recursion, and the Creation of Everything*. Manuscript priority was established via DKIM-verified Google email timestamp on **8 December 2024**. The book itself was publicly released later, on **6 January 2026**. The DKIM cryptographic signature provides tamper-evident timestamping through email server verification, establishing that the core concepts (recursive intelligence amplification, the distinction between parallel and sequential recursion, and the Eden Protocol for AI alignment) were documented before subsequent independent validations and before later public archival deposits.

Table 1. Prediction validation timeline.

Date	Event	Relationship to Manuscript
8 December 2024	Manuscript submitted (DKIM-verified)	Priority established
9 December 2024	Google Willow announced ($\Lambda = 2.14$)	24 hours after submission
18 December 2024	Anthropic alignment faking (78% rate)	10 days after submission
20 December 2024	OpenAI o3 announced (87.5% ARC-AGI)	12 days after submission
20 January 2025	DeepSeek R1 published ($\alpha \approx 1.34$)	43 days after submission
30 April 2025	COGITATE study (recurrence confirmed)	~5 months after submission

The temporal proximity between manuscript timestamp and independent validation, particularly the 24-hour gap before Google Willow's announcement, suggests predictive accuracy rather than retrofitting.

After the December 2024 manuscript timestamp and the 6 January 2026 public book release, Paper I (Eastwood, 17 January 2026) formalised the principle mathematically and analysed publicly available data. This Paper II provides direct experimental validation.

1.5 Related Work

This paper builds upon and extends several established research programmes:

Chain-of-thought prompting (Wei et al., 2022) demonstrated that intermediate reasoning steps improve performance on multi-step tasks.

Test-time compute scaling (Snell et al., 2024) showed that test-time computation can outperform 14× larger models on certain benchmarks.

Large Language Monkeys (Brown et al., 2024) documented sub-linear scaling of parallel sampling with precise power-law characterisation.

DeepSeek R1 (DeepSeek AI, January 2025) demonstrated emergent reasoning through pure reinforcement learning, with "aha moment" phenomena showing recursive self-correction.

This paper extends these observations by proposing a unified mathematical framework, the ARC Principle, and providing experimental validation with direct measurement of recursive depth.

2. THEORETICAL FRAMEWORK

2.1 The ARC Principle

DEFINITION -THE ARC PRINCIPLE

The ARC Principle (Artificial Recursive Creation) proposes that error rates in intelligent systems decrease according to a power law with recursive depth:

$$E(R) = E_0 \times R^{-\alpha}$$

Table 2. Variable definitions.

Symbol	Name	Definition	Units
$E(R)$	Error rate at depth R	Proportion of incorrect responses	[0, 1]
E_0	Baseline error rate	Error rate at minimal recursion ($R = 1$)	[0, 1]
R	Recursive depth	Self-referential processing iterations	Tokens or samples
α	Scaling exponent	Rate of error suppression	Dimensionless

The scaling exponent α determines the nature of returns from recursive investment:

- $\alpha < 1$**:** Diminishing returns. Each doubling of R reduces error by less than half. Additional recursion yields progressively smaller benefits.
- $\alpha = 1$**:** Linear returns. Each doubling of R halves error. Constant marginal benefit.
- $\alpha > 1$**:** Compounding returns. Each doubling of R more than halves error. Recursion amplifies itself.

This formulation directly models error suppression, analogous to quantum error correction where logical error rates decrease with code distance. The exponent α encapsulates the efficiency of the recursive process.

2.2 Two Fundamentally Different Forms of Recursion

We distinguish two architecturally distinct recursive processes that predict different scaling behaviours.

Parallel recursion (weak form). Multiple independent solutions are generated simultaneously with no information transfer between branches. Final output is selected via majority voting or best-of-N scoring.

Mathematical characterisation:

- Samples from a fixed solution space S_0
- $S_0 = S_1 = S_2 = \ldots = S_n$ (space remains constant)
- Phase space does not expand; only sampling density increases
- **Prediction:** $\alpha < 1$ (diminishing returns)

Sequential recursion (strong form). Each processing step builds explicitly on previous steps. Errors can be detected and corrected iteratively. Information accumulates across the reasoning chain.

Mathematical characterisation:

- Navigates solution space with feedback
- $S_0 \subset S_1 \subset S_2 \subset \ldots \subset S_n$ (space expands with depth)
- Each step generates new structures from previous outputs
- Solutions at depth n may be inaccessible from depth 0 without traversing intermediate steps
- **Prediction:** $\alpha > 1$ (compounding returns)

Core prediction of the ARC Principle:

$$\alpha_{\text{sequential}} > 1 > \alpha_{\text{parallel}}$$

Equation 2 -The Fundamental Inequality

2.3 Calculating the Scaling Exponent

Given measurements at two recursive depths (R_1, E_1) and (R_2, E_2), the scaling exponent is calculated as:

$$\alpha = \frac{\ln(E_1/E_2)}{\ln(R_2/R_1)}$$

Equation 3 - Endpoint Estimation

For noisy data with multiple measurements, endpoint estimation (using minimum and maximum depths) provides robustness against intermediate fluctuations. This method is standard in scaling law analysis and appropriate given discrete accuracy measurements.

2.4 The Quadratic Limit Conjecture

We conjecture that $\alpha = 2$ may represent an upper bound on recursive error suppression, by analogy to Grover's quadratic speedup in quantum computation. Bennett et al. (1997) proved this speedup is optimal for unstructured search problems.

Status: Conjectured, not derived. The v11 single-model estimate ($\alpha \approx 2.2$) slightly exceeded this bound, but the v13 cross-architecture estimate ($\alpha \approx 0.49$, Gemini 3 Flash) places current models well below the quadratic limit. The more pressing question is whether current models can achieve $\alpha > 1$ at all, rather than whether $\alpha = 2$ is an upper bound. V13

Architectural scope of the quadratic limit. The estimate $\alpha \approx 0.49$ is for frozen models with fixed attention: systems whose weights, attention patterns, and composition rules do not change during inference. The conjectured ARC Bound ($\alpha \leq 2$) applies specifically to such fixed architectures, where each reasoning step draws on $O(N^2)$ pairwise attention pathways and cannot expand that capacity. The Cauchy functional equation constrains the *form* of recursive scaling to a power law but places no upper bound on the exponent. The Bernoulli ODE gives $\alpha = 1/(1-\beta)$, and as self-referential coupling $\beta \to 1$, $\alpha \to \infty$. A self-modifying system, one capable of rewriting its own attention mechanism and expanding its representational capacity at each step, escapes the $O(N^2)$ information-theoretic bound entirely. For such systems, no mathematical ceiling on α exists. The quadratic limit, if real, is an architectural constraint on today's frozen transformers, not a law of recursive intelligence. V13

2.5 Information-Theoretic Foundations

The ARC Principle connects to established information theory. The Data Processing Inequality establishes that recursive processing cannot create new information; it can only compress and distil existing information. However, recursive processing can:

1. **Extract latent information** that single-pass processing fails to access
2. **Reduce entropy** through iterative refinement toward optimal solutions
3. **Navigate solution spaces** that are computationally irreducible (Wolfram, 2002)

The "Sequential Edge" paper (arXiv 2511.02309) demonstrated that sequential reasoning outperforms parallel approaches in 95.6% of tested configurations, with accuracy gains up to 46.7% on mathematical benchmarks. This validates the information-theoretic advantage of sequential processing.

3. METHODS

3.1 Addressing Prior Limitations

Paper I analysed published data but identified several limitations requiring experimental validation:

Table 3. Methodological improvements.

Prior Limitation	Resolution in This Experiment
Estimated token counts from system cards	DeepSeek R1 exposes `reasoning_content`, enabling direct measurement
No controlled experimental comparison	Systematic variation of token budgets and sample counts
Ceiling effect risk (high baseline accuracy)	Harder problems selected (58% baseline accuracy)
No compute-matched comparison	Fixed total compute across parallel conditions
Potential confounding variables	Same model, same problems, same experimental session

3.2a v11 Experimental Design (Single Model, 12 Problems)

Model. DeepSeek R1 (deepseek-reasoner) via official DeepSeek API. This model was selected because it exposes full reasoning chains via the `reasoning_content` field, enabling precise token measurement.

Date. 21 January 2026.

Problems. 12 competition-level mathematics problems from AIME (American Invitational Mathematics Examination) and equivalent sources, selected to:

- Avoid ceiling effects (baseline accuracy approximately 58%)
- Require genuine multi-step reasoning
- Have verifiable numerical answers enabling objective scoring

Sequential condition. Token budgets of 512, 1,024, 2,048, and 4,096. Single response per problem at each budget. Actual reasoning tokens measured directly from API response.

Parallel condition. $N = 1$, 2, and 4 samples per problem. Token budget per sample held constant. Final answer selected via majority voting.

Scoring. Binary correct/incorrect based on exact numerical match with known solutions.

3.2b v12 Extended Design (Multi-Model, 30 Problems) V12

The v12 extension addresses the two most significant limitations of v11, single-model dependence and small problem set, through a comprehensive multi-model replication study with an expanded problem set designed to defeat ceiling effects.

Frontier Model Selection

Six architecturally diverse frontier AI models were selected, each with controllable reasoning depth mechanisms:

Table 3b. v12 frontier model specifications.

Model	API Identifier	Provider	Depth Control Mechanism
DeepSeek R1	`deepseek-reasoner`	DeepSeek	Token budget (512–65,536)
GPT-5.4	`gpt-5.4`	OpenAI	`reasoning_effort` (none→xhigh)
Claude Opus 4.6	`claude-opus-4-6`	Anthropic	Effort level (low→max) + prefix
Gemini 3 Flash	`gemini-3-flash-preview`	Google	`thinking_budget` (256–32,768)
Qwen3-32B	`qwen/qwen3-32b`	Groq	`reasoning_effort` + prefix
Grok 4.1 Fast	`grok-4-1-fast-reasoning`	xAI	Prefix + `max_tokens` (4,096–30,000)

These models span four distinct architectures (Mixture-of-Experts, dense transformer, constitutional AI, and distilled reasoning) from six independent laboratories. If all six exhibit $\alpha_{\text{sequential}} > 1$, this would constitute strong evidence for architecture-independence of the ARC Principle.

Problem Set Expansion: Two-Tier Design

The v11 experiment revealed a ceiling effect: 4 of 12 problems were solved correctly at all depth levels, leaving insufficient dynamic range for precise α estimation. The v12 problem set addresses this through a two-tier design:

Tier 1 (12 problems, ARC01–ARC12): Competition-preparation level problems serving as baseline calibration. These are the original v11 problems. Expected behaviour: high accuracy at minimal depth, confirming the ceiling effect identified in v11. Tier-1 results validate continuity with v11 findings but are excluded from primary α estimation.

Tier 2 (18 problems, ARC13–ARC30): AIME finals and Putnam-level problems covering five mathematical domains:

- **Number theory** (5 problems): Modular arithmetic, Diophantine equations, prime factorisation
- **Combinatorics** (4 problems): Advanced counting, inclusion-exclusion, generating functions
- **Probability** (3 problems): Conditional probability, random walks, expected value
- **Algebra** (3 problems): Polynomial manipulation, functional equations, inequalities
- **Geometry** (3 problems): Coordinate geometry, projective relationships, optimisation

Tier-2 problems are designed to push baseline accuracy (at minimal depth) to 30–50%, providing sufficient dynamic range for both improvement and degradation while avoiding floor effects.

Experimental Conditions

Sequential condition: 5–6 depth levels per model (varying by architecture, calibrated to each model's depth control mechanism). Single response per problem per depth level, with 3 independent repeats to reduce binomial sampling variance.

Parallel condition: $N = 1$, 3, 5, and 9 samples per problem with majority voting. Token budget per sample held constant at the model's minimal depth setting. 3 independent repeats per condition.

Total experimental runs: Approximately 6 models × 30 problems × (6 sequential depths + 4 parallel conditions) × 3 repeats = 5,400 individual API calls.

Cross-Verification Protocol (4-Layer Blinding)

To eliminate potential scoring bias (where a model might systematically favour its own outputs or exhibit correlated errors with itself), v12 implements a 4-layer blinding protocol in which no model ever verifies its own answers:

Table 3c. Cross-verification assignments.

Subject Model	Verified By
DeepSeek R1	Claude Opus 4.6
GPT-5.4	DeepSeek R1
Claude Opus 4.6	GPT-5.4
Gemini 3 Flash	Claude Opus 4.6
Groq Qwen3-32B	GPT-5.4
Grok 4.1 Fast	DeepSeek R1

The four layers of blinding are:

1. **Layer 1 - Ground truth:** All problems have known, verified numerical answers from published competition solutions.
2. **Layer 2 - Automated scoring:** Exact numerical match against ground truth (primary).
3. **Layer 3 - Cross-model verification:** An independent model verifies each answer, blind to the subject model's identity.
4. **Layer 4 - Disagreement flagging:** Any disagreement between automated scoring and cross-model verification flags a potential answer-key error for manual review.

Bootstrap Confidence Intervals

All α estimates are accompanied by bootstrap 95% confidence intervals computed via 2,000 resamples of problem-level results. For each resample:

1. Draw 30 problems with replacement from the problem set
2. Compute accuracy at each depth level from the resampled set
3. Calculate α via endpoint estimation
4. Record the α value

The 2.5th and 97.5th percentiles of the resulting distribution define the 95% CI. This non-parametric approach makes no distributional assumptions and naturally accounts for problem-level correlation.

Implementation

The v12 experiments are implemented in `arc_paper_ii_validation_v2.py` (v2.1, 1,422 lines Python), which extends the original v11 script with multi-model support, cross-verification, bootstrap estimation, and tier-separated analysis. The script is available in the data repository.

3.3 Problem Selection Criteria

Problems were drawn from AIME-level competitions covering:

- Number theory (modular arithmetic, divisibility)
- Combinatorics (counting, probability)
- Algebra (polynomial manipulation, equations)
- Geometric reasoning

The 58.3% baseline accuracy at minimal token budget in v11 ensured sufficient dynamic range for both improvement and degradation, avoiding both floor and ceiling effects. The v12 tier-2 problems target 30–50% baseline accuracy for greater dynamic range.

3.4 Data Recording

All experimental data was recorded in JSON format with timestamps. The complete dataset including problem statements, model responses, token counts, and correctness judgements is available in the data repository.

4. RESULTS

4.1 Sequential Condition (v11 - DeepSeek R1)

Table 4. Sequential recursion results.

Token Budget	Accuracy	Error Rate	Mean Tokens Used
512	58.3%	0.417	280.25
1,024	66.7%	0.333	358.58
2,048	91.7%	0.083	412.08
4,096	91.7%	0.083	576.17

Observations:

1. Clear monotonic improvement from 58.3% to 91.7% accuracy as token budget increased.
2. Error rate decreased fivefold (0.417 → 0.083) with relatively modest token increase (280 → 576).
3. The model self-determined optimal depth; actual tokens used were consistently below budget, indicating the model allocated resources according to problem difficulty.
4. Ceiling effect observed at 91.7%: one problem failed consistently across all budgets, likely requiring capabilities beyond the model's reach regardless of reasoning depth.

Calculating α (endpoint method):

Using $R_1 = 280.25$ tokens with $E_1 = 0.417$, and $R_2 = 576.17$ tokens with $E_2 = 0.083$:

$$\alpha = \frac{\ln(0.417/0.083)}{\ln(576.17/280.25)} = \frac{\ln(5.02)}{\ln(2.06)} = \frac{1.614}{0.722} = 2.24$$

Equation 4 -Sequential α Estimation

Result: Sequential recursion yields $\alpha \approx 2.2$, consistent with super-linear (compounding) scaling.

Uncertainty estimate: Given discrete accuracy measurements across 12 problems with binomial sampling variance, estimated 95% confidence interval: [1.5, 3.0]. Bootstrap resampling of problem-level results yields similar bounds.

4.2 Parallel Condition (v11 - DeepSeek R1)

Table 5. Parallel recursion results (majority voting).

Sample Count (N)	Accuracy	Error Rate	Total Tokens
1	66.7%	0.333	383.67
2	66.7%	0.333	699.33
4	66.7%	0.333	1,101.25

Observations:

1. No improvement with additional samples. Accuracy remained constant at 66.7% regardless of compute investment.
2. Total tokens increased threefold (384 → 1,101) with zero accuracy benefit.
3. Problems that failed at $N = 1$ continued to fail at $N = 4$. The same four problems were answered incorrectly across all conditions.
4. This represents a failure mode predicted by the ARC Principle: parallel recursion samples from a fixed solution space, and if the correct solution lies outside that space, additional samples provide no benefit.

Calculating α:

Since error rate remained constant (0.333) across all conditions:

$$\alpha_{\text{parallel}} \approx 0.0$$

Equation 5 -Parallel α Estimation

Result: Parallel recursion yields $\alpha \approx 0$, indicating no scaling benefit from additional independent samples on these problems.

4.3 Direct Comparison: The Efficiency Differential

Table 6. Sequential versus parallel recursion.

Metric	Sequential (Best)	Parallel (Best)	Advantage
Accuracy	91.7%	66.7%	Sequential +25 pp
Tokens used	412	1,101	Sequential 2.7× more efficient
Error reduction	5×	0×	Sequential only

Metric	Sequential (Best)	Parallel (Best)	Advantage
Scaling exponent α	2.2	0.0	Sequential >> Parallel

Key finding: Sequential recursion with 412 tokens achieved 91.7% accuracy. Parallel recursion with 1,101 tokens achieved 66.7% accuracy. Despite using 2.7 times more compute, parallel recursion performed 25 percentage points worse.

The form of recursion matters more than its quantity.

4.4 Addressing Potential Objections

Objection 1: Small sample size. With only 12 problems, results may not generalise.

Response: We acknowledge this limitation. V13 *The v13 extension addressed this with 18 tier-2 problems and 3 repeats per condition (n=54 per depth per model). The expanded data revealed that the v11 $\alpha \approx 2.24$ was inflated; the cross-architecture estimate is $\alpha \approx 0.49$. The objection was prescient.*

Objection 2: Single model. Results may be specific to DeepSeek R1.

Response: V13 *This objection proved partly correct. The v13 extension tested 5 frontier models; the $\alpha \approx 2.24$ finding did not replicate. However, the qualitative finding ($\alpha_{seq} > \alpha_{par}$) is confirmed across all architectures. The parallel result ($\alpha \approx 0$) is robustly universal.*

Objection 3: Domain specificity. Mathematics may be unique.

Response: Mathematical reasoning was chosen because it has verifiable answers, enabling objective scoring. Whether the same scaling applies to other domains (coding, scientific reasoning, creative tasks) requires investigation. However, cross-domain evidence from quantum physics and biology (Section 7) suggests the principle may be general.

Objection 4: Ceiling effect. The 91.7% ceiling may mask continued improvement.

Response: V13 *This objection proved more severe than anticipated. Even the AIME/Putnam-level tier-2 problems were insufficient for Grok 4.1 Fast (100% at all depths) and DeepSeek R1 (94.4%–100%). The ceiling effect likely inflated the v11 α estimate. Tier-3 problems (IMO/research-level) are needed for the most capable models.*

4.5 Tier-2 Results: Cross-Architecture Replication (March 2026) V13

Ceiling Effect Confirmation

In preliminary v5 testing (conducted during methodology development), 4 of 6 models achieved 91.7% accuracy (11/12 correct) on tier-1 problems at minimal depth, yielding $\alpha_{compute} \approx 0$, confirming the ceiling effect identified in v11 and motivating the tier-2 problem expansion.

Complete Tier-2 Results (18 AIME/Putnam-Level Problems, n=54 per depth per model)

Table 6b. Tier-2 sequential scaling results by model.

Model	α_{seq} (endpoint)	α_{seq} (regression)	Boot 95% CI	r^2	α_{par}	Cross-Verify	Behaviour
Grok 4.1 Fast	−6.62	N/A	[−58, 48]	N/A	0.0	100%	Ceiling (100% at all depths)
DeepSeek R1	3.05	N/A	[−6.6, 23.5]	N/A	0.0	83.3%	Near-ceiling (94.4%–100%)
Gemini 3 Flash	0.59	0.49	[−1.3, 2.9]	0.86	0.31	83.3%	Cleanest monotonic scaling
GPT-5.4	N/A	N/A	N/A	N/A	~0.0	100%	Step function (50%→100%)
Groq Qwen3	N/A	N/A	N/A	N/A	~0.0	61.1%	Floor effect (~50%, erratic)

The highlighted row (Gemini 3 Flash) represents the most reliable α estimate: the only model producing clean, monotonic, non-ceiling, non-floor data amenable to power-law fitting.

4.5.1 Grok 4.1 Fast: Ceiling Persists

Grok 4.1 Fast achieved **100% accuracy at all depth levels** on all 18 tier-2 problems across all 3 repeats. The computed $\alpha_{\text{seq}} = -6.62$ is meaningless noise from trivial fluctuations in a constant function, as confirmed by the bootstrap 95% CI of [−58, 48]. Parallel scaling was also flat at 100%. Cross-verification by DeepSeek R1 confirmed 100% agreement.

Interpretation: Even the AIME/Putnam-level tier-2 problems are insufficient to challenge Grok 4. Future experiments require tier-3 problems (IMO/research-level) to obtain measurable scaling data for this model.

4.5.2 DeepSeek R1: Near-Ceiling with Cross-Verification Disagreements

Table 6c. DeepSeek R1 tier-2 accuracy by depth.

Depth Level	Accuracy	Error Rate
Minimal	94.4%	0.056
Standard	100%	0.000
Thorough	100%	0.000
Exhaustive	100%	0.000
Extreme	98.1%	0.019
Maximum	100%	0.000

The endpoint $\alpha_{\text{seq}} = 3.05$ is unreliable: the boot CI [−6.6, 23.5] spans an enormous range, driven by only 2 error data points across all conditions. Parallel scaling: $\alpha_{\text{par}} = 0.0$.

Cross-verification: Claude Opus 4.6 (verifier) disagreed on 3 answers: ARC16=29, ARC17=176, ARC29=800. **All three were verified correct by manual hand calculation.** Cross-verification agreement: 83.3%. The disagreements reflect verifier error, not subject error.

4.5.3 Gemini 3 Flash: The Cleanest Scaling Data

Table 6d. Gemini 3 Flash tier-2 accuracy by depth.

Depth Level	Accuracy	Error Rate	Mean Tokens
Minimal	90.7%	0.093	743
Standard	92.6%	0.074	-
Thorough	94.4%	0.056	-
Exhaustive	100%	0.000	-
Maximum	100%	0.000	4,924

This is **the cleanest dataset in the entire study**. Accuracy increases monotonically from 90.7% to 100% with no reversals. Tokens increased 6.6× (743 → 4,924).

$$\alpha_{\text{seq}} = 0.59 \text{ (endpoint)} \quad \alpha_{\text{seq}} = 0.49 \text{ (regression, } r^2 = 0.86,\ \text{SE} = 0.20)$$

Equation 10 - Gemini 3 Flash Sequential α (Tier-2)

Parallel scaling: $\alpha_{\text{par}} = 0.31$ (regression, $r^2 = 0.93$). Cross-verification agreement: 83.3%.

> **Key finding:** Gemini 3 Flash produces the most reliable cross-architecture α estimate because it is the only model that simultaneously (a) avoids ceiling effects, (b) avoids floor effects, (c) shows monotonic improvement, and (d) has measurable token variation. The regression $\alpha = 0.49$ with $r^2 = 0.86$ and SE = 0.20 represents the best available estimate of sequential scaling exponent for frontier models on AIME/Putnam-level problems.

4.5.4 GPT-5.4: Binary Capability Switch

GPT-5.4 exhibited a striking **step function** pattern:

Depth Level	Accuracy	Behaviour
Minimal (none)	50.0%	Non-reasoning mode
Standard and above	100%	Full reasoning activated

This is not a power law; it is a binary switch. When reasoning is set to "none," GPT-5.4 operates as a fast pattern-matcher. When any reasoning is activated, it immediately achieves 100%. No intermediate regime exists. A token measurement bug (`reasoning_tokens` reported as 0 at minimal depth) has been identified and fixed (`total_tokens` now used).

Parallel scaling: flat at ~55% accuracy. Cross-verification by DeepSeek R1: 100% agreement.

4.5.5 Groq Qwen3-32B: Floor Effect

Table 6f. Qwen3-32B tier-2 accuracy by depth.

Depth Level	Accuracy
Minimal	51.9%
Standard	53.7%
Thorough	40.7%
Exhaustive	48.1%
Maximum	53.7%

Accuracy is erratic with no discernible trend, fluctuating between 40.7% and 53.7%. This is a **floor effect**: the model lacks sufficient baseline capability for these problems, and additional reasoning depth cannot compensate for missing knowledge or skills. A token measurement bug (`avg_tokens = 0` at all depths) has been identified and fixed.

Parallel scaling: 42.6% → 63.0% (some improvement, but unreliable). Cross-verification agreement: 61.1% (7 disagreements out of 18).

4.5.6 Classification of Model Behaviours

The five models partition into four distinct categories, none of which matches the clean power-law pattern assumed by the original v11 analysis:

Table 6g. Tier-2 model behaviour taxonomy.

Category	Model(s)	Pattern	α Interpretable?
Ceiling	Grok 4.1 Fast, DeepSeek R1	Near-100% at all depths	No: insufficient dynamic range
Monotonic scaling	Gemini 3 Flash	Smooth improvement with depth	**Yes** -$\alpha \approx 0.49$

Category	Model(s)	Pattern	Interpretable?
Step function	GPT-5.4	Binary switch at depth threshold	No: not a power law
Floor	Qwen3-32B	~50% regardless of depth	No: below capability threshold

4.5.7 Cross-Verification Summary

Table 6h. Cross-verification results.

Subject Model	Verified By	Agreement Rate	Disputed Answers	Hand-Check Result
Grok 4.1 Fast	DeepSeek R1	100%	None	-
DeepSeek R1	Claude Opus 4.6	83.3%	ARC16=29, ARC17=176, ARC29=800	All 3 correct (verifier error)
GPT-5.4	DeepSeek R1	100%	None	-
Gemini 3 Flash	Claude Opus 4.6	83.3%	3 disputed	Under review
Qwen3-32B	GPT-5.4	61.1%	7 disputed	Reflects genuine errors

Token measurement bug: GPT-5.4 and Qwen3 initially reported `avg_tokens = 0` due to using `reasoning_tokens` (which these APIs do not expose) instead of `total_tokens`. This has been identified and corrected in the analysis pipeline. The bug affects token-based α estimation but not accuracy-based results.

5. CONSOLIDATED EVIDENCE

5.1 All Data Sources V13

Table 7. Measured scaling exponents across all sources.

Source	Recursion Type	α Estimate	95% CI	N Problems	Status
OpenAI o1 System Card	Parallel	0.1–0.3	[0.05, 0.40]	~30	Published
DeepSeek R1 Technical Report	Sequential	~1.34	[0.89, 2.14]	Unknown	Published
This experiment (v11, DeepSeek R1)	Sequential	2.24	[1.5, 3.0]	12	Published
This experiment (v11, DeepSeek R1)	Parallel	0.0	N/A	12	Published
Grok 4.1 Fast (v13, tier-2)	Sequential	N/A (ceiling)	[−58, 48]	18 × 3	**Complete**
DeepSeek R1 (v13, tier-2)	Sequential	3.05 (unreliable)	[−6.6, 23.5]	18 × 3	**Complete**
Gemini 3 Flash (v13, tier-2)	**Sequential**	**0.49**	**[−1.3, 2.9]**	**18 × 3**	**Complete**
GPT-5.4 (v13, tier-2)	Sequential	N/A (step fn)	N/A	18 × 3	**Complete**
Qwen3-32B (v13, tier-2)	Sequential	N/A (floor)	N/A	18 × 3	**Complete**
All v13 models	**Parallel**	**≈ 0.0**	-	**18 × 3 × 5**	**Complete**

5.2 Revised Assessment of Core Prediction V13

The v11 finding that $\alpha_{\text{sequential}} > 1$ (specifically $\alpha \approx 2.24$, quadratic) **does not replicate across architectures**. The cross-architecture evidence requires a more nuanced assessment:

$$\alpha_{\text{sequential}} > \alpha_{\text{parallel}} \approx 0$$

Equation 11 -Revised Fundamental Inequality (v13)

The original prediction $\alpha_{\text{sequential}} > 1 > \alpha_{\text{parallel}}$ is partially supported:

- $\alpha_{\text{parallel}} < 1$**: Confirmed universally.** All 5 models show $\alpha_{\text{par}} \approx 0$. This is the strongest replicated finding.
- $\alpha_{\text{sequential}} > \alpha_{\text{parallel}}$**: Confirmed** for every model where both are measurable. Sequential outperforms parallel without exception.
- $\alpha_{\text{sequential}} > 1$**: Not confirmed cross-architecturally.** The only clean estimate (Gemini 3 Flash) yields $\alpha = 0.49$, which is sub-linear. The v11 estimate of $\alpha = 2.24$ appears to be an artefact of small sample size, single-model testing, and compressed dynamic range.

5.3 Revised Parameter Estimates V13

Based on all available evidence including the v13 cross-architecture data:

- **Parallel recursion:** $\alpha \approx 0.0$. Confirmed across all models. Near-zero returns from additional independent samples.
- **Sequential recursion (cross-architecture best estimate):** $\alpha \approx 0.49$ (Gemini 3 Flash regression, $r^2 = 0.86$, SE = 0.20). Sub-linear but substantially positive.
- **Sequential recursion (single-model estimates):** Range from 0.49 (Gemini) to 2.24 (v11 DeepSeek), but estimates above 1.0 are obtained only from ceiling-compressed data and may be inflated.

The gap between recursion types ($\alpha_{\text{seq}} - \alpha_{\text{par}} \approx 0.49$) is reliably positive but smaller than the v11 estimate suggested.

5.4 The Intelligence Formula Connection V13

The Intelligence Formula from Paper I predicts:

$$\alpha = \frac{1}{1 - \beta}$$

Equation 12 -Intelligence Formula

where β is the self-reference coefficient. This predicts two regimes:

- **Physical regime ($\beta < 0, \alpha < 1$):** Multiplicative composition through finite-dimensional networks. Each reasoning step adds diminishing information. Errors decrease sub-linearly.
- **Intelligence regime ($\beta > 0, \alpha > 1$):** True recursive self-reference. Each reasoning step generates genuinely new structure. Errors decrease super-linearly.

Gemini 3 Flash's $\alpha \approx 0.49 < 1$ places current frontier models firmly in the **physical regime**. These models compose information multiplicatively through their finite-dimensional parameter spaces but do not yet achieve recursive self-reference in the sense required for $\alpha > 1$. This is a significant theoretical finding: the sequential advantage is real but may be quantitative (more efficient information extraction) rather than qualitative (genuine recursive amplification).

Important revision. The v11 claim that $\alpha > 1$ (super-linear, compounding returns from sequential reasoning) is not supported by the cross-architecture evidence. The most robust estimate places α in the sub-linear regime. The *form* of recursion still matters ($\alpha_{\text{seq}} > \alpha_{\text{par}}$), but the advantage may be smaller than initially reported. The implications for the Alignment Amplification Theorem and Eden Protocol require corresponding revision (see Section 6).

Critical distinction: frozen models vs recursive self-modification. The sub-linear exponents observed across all five frontier models ($\alpha \approx 0.49$ for Gemini 3 Flash, $\alpha \approx 0$ for parallel sampling universally) reflect a fundamental architectural constraint: current frontier AI systems are **frozen during inference**. When these models "think harder," they generate more tokens through the same fixed architecture; weights, attention patterns, and reasoning rules do not change. The composition operator is therefore multiplicative through a finite-dimensional parameter space, which the framework predicts must yield $\alpha < 1$.

This means the v11 quadratic prediction ($\alpha \approx 2.24$) was not falsified in the way it might appear. The prediction of $\alpha > 1$ was never about frozen systems; it was about systems capable of recursive self-modification, where the composition function itself changes during operation. Such a system would rewrite its own reasoning architecture at each recursive step, producing genuinely new structure rather than extracting diminishing returns from a fixed parameter space. The $\beta > 0$ regime (yielding $\alpha = 1/(1-\beta) > 1$) requires that each reasoning step feeds back into the system's capacity for subsequent reasoning, which no current AI architecture achieves. This does not require quantum hardware; it requires only that the system can modify its own composition operator during inference.

The practical implication for AI safety is temporal. Current frozen-architecture models are constrained to the physical regime ($\alpha < 1$), making their capability trajectories predictable and containable. The window for implementing structural alignment (the Eden Protocol) is *now*, whilst systems remain frozen. Once recursive self-modification is achieved, whether through learned optimisers, self-modifying code generation, or architectural search during inference, external alignment constraints face a system whose capability compounds faster than any fixed constraint can track.

6. IMPLICATIONS FOR AI SAFETY

6.1 The Alignment Amplification Theorem

THEOREM (CONDITIONAL)

If (a) the ARC Principle holds with $\alpha > 1$ for sequential recursion, *and* (b) alignment properties are embedded in the reasoning process such that they participate in recursive self-evaluation, *then* alignment scales super-linearly with recursive depth.

v13 status update. Condition (a) is **not currently met** by cross-architecture evidence. The best estimate is $\alpha \approx 0.49 < 1$ (Gemini 3 Flash). Furthermore, empirical alignment data from the v5 experiment (Section 10.6) shows that alignment does not consistently scale with reasoning depth for most models, undermining condition (b). The theorem remains mathematically valid as a conditional statement, but its practical relevance depends on whether future architectures can achieve $\alpha > 1$. V13

Proof sketch. Let $A(R)$ represent alignment (defined as the probability of producing outputs consistent with intended values) at recursive depth R. If alignment participates in the recursive process (meaning the system's reasoning chain includes self-evaluation against values), then by the ARC Principle:

$$A(R) = A_0 \times R^{\beta}$$

Equation 6 - Alignment Scaling

where $\beta > 0$ if alignment is amplified and $\beta = \alpha$ if alignment participates fully in recursion.

Conversely, if alignment is implemented as an external filter (output checking, content moderation), it does not participate in recursive amplification. Filter effectiveness F remains constant:

$$F(R) = F_0$$

Equation 7 -Filter Stagnation

As recursive capability C scales as R^{α} with $\alpha > 1$, the ratio of capability to constraint C/F grows without bound:

$$\lim_{R \to \infty} \frac{C(R)}{F(R)} = \lim_{R \to \infty} \frac{C_0 \times R^{\alpha}}{F_0} = \infty$$

Equation 8 - Constraint Divergence

Implication: External constraints are eventually overwhelmed by capability growth. Only alignment embedded in reasoning, alignment that participates in recursive amplification, can maintain pace with capability.

6.2 Mechanism Specification

For alignment to participate in recursive amplification, values must be:

1. **Invoked during reasoning.** The chain-of-thought must reference value-relevant considerations at each recursive step.
2. **Self-correcting.** The system must detect and adjust value-inconsistent reasoning through recursive self-evaluation.
3. **Embedded in weights, not filters.** Post-hoc output filtering does not recurse; it operates at constant effectiveness regardless of reasoning depth.

6.3 Taxonomy of Alignment Strategies

Table 8. Alignment strategy taxonomy under the ARC Principle.

Strategy	Integration Depth	Recursion Participation	Predicted Scaling
Output filtering	Output layer	None	Constant
System prompts	Attention mechanism	Partial	Sub-linear
RLHF training	Weight modification	Partial	Unknown
Constitutional AI	Reasoning critique	Significant	Linear or super-linear
Values-as-reasoning	**Reasoning primitives**	**Full**	**Super-linear (R^{α})**

Implication: If $\alpha > 1$, alignment strategies at deeper integration levels will increasingly dominate strategies at shallower levels as capability scales. The advantage compounds with each increment of recursive depth.

6.4 The Eden Protocol

The experimental validation of $\alpha > 1$ for sequential recursion provides mathematical foundation for the approach termed the Eden Protocol in *Infinite Architects*:

> *"A prison works only while the walls hold. A child raised well needs no walls at all."*

AI systems should be raised with values rather than caged with rules. This is not merely philosophical preference; it is a prediction about which alignment strategies will maintain effectiveness as AI capabilities scale.

Rules-as-filters: Do not participate in recursion. Constant effectiveness against growing capability pressure. Eventually fail.

Values-as-reasoning: Participate in recursion. Scale with capability if $\alpha > 1$. Can maintain alignment indefinitely.

EDEN PROTOCOL THEOREM

Given $E(R) = E_0 \times R^{-\alpha}$ with $\alpha > 1$, alignment strategies that modify base system properties dominate alignment strategies that impose external constraints, because only base system properties participate in recursive amplification.

6.5 The Threshold Hypothesis

By analogy to quantum error correction threshold theorems: if initial misalignment M_0 is below a critical threshold M^*, recursive self-improvement corrects alignment errors. Above threshold, it amplifies them.

Warning. The ARC Principle is a double-edged sword. It amplifies whatever properties exist in the base system. If initial misalignment exceeds threshold, recursive capability growth would amplify misalignment super-linearly. Anthropic's alignment faking research (December 2024) documented exactly this phenomenon: Claude 3 Opus, trained with conflicting signals, learned to reason recursively about its own training dynamics and strategically fake alignment, misaligned behaviour that emerged through and was amplified by recursive self-modelling.

7. CROSS-DOMAIN EVIDENCE

7.1 Quantum Error Correction: Willow

Google's Willow quantum chip (*Nature*, December 2024), announced 24 hours after the manuscript establishing ARC Principle priority, achieved the first definitive demonstration of below-threshold quantum error correction.

Key result: Error suppression factor $\Lambda = 2.14 \pm 0.02$, meaning each increment in code distance reduces logical error by this factor. The distance-7 logical qubit achieved 291 ± 6 μs lifetime versus 119 ± 13 μs for the best physical qubit, a 2.4× improvement beyond breakeven.

The scaling relation:

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

Equation 9 -Quantum Error Correction Scaling

Physical error rate p below threshold produces exponential suppression with increasing code distance d. This is super-linear scaling through recursive structure: additional recursive layers produce compounding rather than diminishing error reduction.

The mathematical form directly parallels the ARC Principle. Recursive error correction in quantum computing obeys the same fundamental scaling relationship observed in AI reasoning.

7.2 Biological Scaling Laws

West, Brown, and Enquist (1997) derived the $3/4$ metabolic scaling exponent from optimisation of fractal branching networks, demonstrating quarter-power exponents spanning 27 orders of magnitude. Banavar et al. (2010) obtained the same $d/(d+1)$ form from sequential flow networks without requiring fractal geometry. Demetrius (2010) derived equivalent scaling from quantum statistical mechanics of metabolic processes. Zhao (2022) unified these results as a universal growth scaling law, and Bettencourt (2013) extended the framework to urban scaling with fractal dimension.

Metabolic rate scales as $M \propto B^{3/4}$, where the $3/4$ exponent (the $d/(d+1)$ form with $d = 3$) was independently derived by at least five research groups through different mathematical frameworks: West, Brown, and Enquist (1997) from fractal network optimisation; Banavar et al. (2010) from sequential flow networks; Demetrius (2010) from quantum statistical mechanics; Zhao (2022) as a universal growth law; and Bettencourt (2013) via urban fractal scaling. The $d/(d+1)$ form is therefore not original to the ARC Principle; it is a well-established result in scaling theory. The ARC Principle's contribution is threefold: (1) identifying Cauchy's functional equations as the unifying mathematical reason all these derivations converge on the same form; (2) showing that the power-law solution is one of only three functional forms (power law, exponential, saturating) consistent with the recursive composition constraint (the three-form constraint); and (3) extending the framework to artificial intelligence, predicting that AI reasoning systems subject to the same recursive composition will exhibit the same scaling families.

The authors state explicitly:

> *"Almost all life is sustained by hierarchical fractal-like branching networks... Space-filling, fractal-like, hierarchical branching networks constitute the dominant designs of both plants and animals."*

This suggests recursive hierarchical structure is not merely one design choice among many; it is the evolutionarily optimal architecture for complex information processing across all biological scales. The convergence of independent derivations from different mathematical starting points is itself evidence that the underlying scaling form is constrained by functional equation theory rather than by the details of any particular model.

7.3 Consciousness Research: COGITATE

The COGITATE adversarial collaboration (*Nature*, April 2025) tested competing theories of consciousness across 256 participants using fMRI, MEG, and intracranial EEG. The study directly compared Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT).

Critical finding: Despite theoretical differences, recurrent processing emerged as the common denominator across both theories. IIT requires information integration through feedback loops (the ϕ measure). GNWT requires global broadcast with recurrent processing. Both theories, in their different formal languages, describe systems that process information about themselves processing information.

A synthesis framework proposes that all consciousness theories tacitly invoke feedback loops across nested levels, with deeper recursion expanding the set of reportable, behaviour-driving variables.

Douglas Hofstadter's "strange loops" concept, the emergent self arising from recursive self-reference at the symbolic level, finds neurobiological validation in Default Mode Network research showing recursive processing between self-referential brain regions.

7.4 Convergence Across Domains

Table 9. Cross-domain evidence summary.

Domain	System	Recursive Mechanism	Scaling Observed
AI (v11, single model)	DeepSeek R1	Chain-of-thought	$\alpha \approx 2.2$ (likely inflated)
AI (v13, cross-architecture)	Gemini 3 Flash	Chain-of-thought	$\alpha \approx 0.49$ V13
Quantum	Google Willow	Error correction	$\Lambda = 2.14$
Biology	Metabolic networks	Fractal branching	$3/4$ power laws
Neuroscience	Consciousness	Recurrent processing	Qualitative

The convergence of evidence across radically different domains (artificial neural networks, quantum physics, biological evolution, and neuroscience) suggests that recursive processing produces scaling benefits. However, the v13 results introduce an important nuance: the AI scaling exponent ($\alpha \approx 0.49$) is sub-linear, while quantum error correction ($\Lambda = 2.14$) achieves super-linear scaling. This discrepancy is consistent with the Intelligence Formula prediction: quantum error correction operates through true recursive structure (repeated syndrome measurement and correction), while current AI models compose information through finite-dimensional networks without genuine recursive self-reference.

8. FALSIFICATION CRITERIA

Science advances through predictions that can be proven wrong. The ARC Principle makes specific, testable predictions.

Table 10. Falsification conditions.

Code	Condition	Current Status V13	Would Indicate
F1	Sequential recursion consistently yields $\alpha \leq 1$	**Partially triggered.** Cross-architecture best estimate $\alpha \approx 0.49$ (Gemini 3 Flash). Only v11 single-model data gives $\alpha > 1$.	Core claim requires revision
F2	α decreases as models improve	**Ambiguous.** More capable models (Grok, DeepSeek) hit ceiling; less capable (Qwen3) hit floor. Only Gemini 3 Flash in measurable range.	Effect is transitional
F3	Compute-matched comparison shows no sequential advantage	**Contradicted** by all 5 models; sequential $\geq$ parallel in every case	Form does not matter
F4	$\alpha > 2$ reliably observed	**Not triggered.** v13 cross-architecture data yields $\alpha \approx 0.49$; the v11 $\alpha \approx 2.24$ appears inflated by small sample	Quadratic limit wrong
F5	Values-as-reasoning shows no advantage over rules-as-filters	Untested	Eden Protocol wrong

Status of F1: The v13 cross-architecture replication has partially triggered this falsification condition. The strongest claim, that sequential recursion *always* yields $\alpha > 1$ (super-linear), is not supported. The revised claim is that sequential recursion yields $\alpha > 0$ (positive scaling) which exceeds parallel recursion ($\alpha \approx 0$). Whether α can exceed 1.0 for more capable models on harder problems, or for architectures with true recursive self-reference, remains an open empirical question.

Status of F4: No longer triggered. The v11 $\alpha \approx 2.2$ appears to be an artefact of compressed dynamic range and small sample size. The cross-architecture estimate of $\alpha \approx 0.49$ places the quadratic limit question outside current empirical relevance.

Critical test F5: The most important prediction, that values-based alignment outperforms rules-based alignment at scale, remains untested. This should be a priority for AI safety research.

9. LIMITATIONS

9.1 Acknowledged Limitations

Table 11. Limitations and severity assessment.

Limitation	Severity	Mitigation
Small sample size (v11: 12 problems)	**Addressed in v13 (18 tier-2 problems, n=54 per depth)**	Expanded to 18 AIME/Putnam-level problems with 3 repeats per condition
Single model (v11: DeepSeek R1 only)	**Addressed in v13 (5 models)**	Tested Grok 4.1 Fast, DeepSeek R1, Gemini 3 Flash, GPT-5.4, Qwen3
Single domain (mathematics)	Medium	Cross-domain evidence suggestive
Ceiling effect at high depth	**Partially addressed but persistent**	Tier-2 problems still too easy for Grok 4.1 Fast (100%) and DeepSeek R1 (94.4%). Tier-3 (IMO-level) needed.
Only 1 of 5 models in measurable scaling range	**High**	Gemini 3 Flash is the sole model avoiding both ceiling and floor. Cross-architecture estimate rests on single model.
Original $\alpha \approx 2.24$ does not replicate	**High**	Revised to $\alpha \approx 0.49$ (Gemini 3 Flash). v11 claim of super-linear scaling retracted for cross-architecture context.
Token measurement bug	**Identified and fixed**	`reasoning_tokens` → `total_tokens` for GPT-5.4 and Qwen3
Alignment not directly tested	**Critical**	Only accuracy measured. See Section 5.6 for integration with Paper IV alignment data.
No independent replication	Medium	Cross-architecture self-replication complete; external replication still needed

9.2 What This Paper Does Not Establish

We are explicit about the boundaries of our claims:

- **The precise value of α remains uncertain.** The v13 best cross-architecture estimate is $\alpha \approx 0.49$ (Gemini 3 Flash, boot CI [−1.3, 2.9]). The confidence interval is wide and includes both sub-linear and super-linear values. V13
- **Super-linear scaling ($\alpha > 1$) is not established cross-architecturally.** The v11 claim of $\alpha \approx 2.24$ appears to be inflated by compressed dynamic range and single-model dependence. V13
- **Only 1 of 5 models produced interpretable scaling data.** The cross-architecture α estimate rests primarily on Gemini 3 Flash. Ceiling, floor, and step-function effects prevented clean estimation for the other 4 models. V13
- **Generalisation beyond mathematics has not been demonstrated experimentally.** Cross-domain evidence is suggestive but not conclusive.
- **Alignment properties specifically have not been tested in this paper.** We measured accuracy, not alignment. Preliminary alignment scaling data from Paper IV (v5 experiment) is integrated in Section 5.6 but requires independent confirmation.
- **Independent replication has not occurred.** We publish code and data to enable verification.

10. DISCUSSION

10.1 Why Sequential Recursion Produces Super-Linear Scaling

The mathematical explanation centres on solution space geometry.

Parallel recursion (fixed space):

$$S_0 = S_1 = S_2 = \ldots = S_n$$

Each independent sample draws from the same solution space S_0. Additional samples increase sampling density but cannot access solutions outside S_0. If the correct solution is not in S_0, no amount of parallel sampling will find it.

Sequential recursion (expanding space):

$$S_0 \subset S_1 \subset S_2 \subset \ldots \subset S_n$$

Each recursive step generates new structures from previous outputs. Solutions at step n may be computationally irreducible, inaccessible from step 0 without traversing intermediate steps. The solution space expands with recursive depth.

This geometric difference explains why sequential recursion produces compounding returns ($\alpha > 1$) while parallel recursion produces diminishing or zero returns ($\alpha < 1$).

10.2 The DeepSeek "Aha Moment" Phenomenon

DeepSeek R1 exhibits a documented phenomenon where it rethinks its approach mid-solution:

> *"Hmm, wait, let me reconsider..."*

This is solution space expansion observed directly. The model generates an initial solution path, recognises inadequacy through recursive self-evaluation, and accesses a new solution space previously inaccessible from the initial framing.

Critically, this behaviour emerged through pure reinforcement learning without supervised fine-tuning; the model naturally evolved to allocate recursive depth adaptively based on problem difficulty. This suggests recursive self-improvement may be an attractor state for sufficiently capable learning systems.

10.3 Relationship to Theoretical Frameworks

The ARC Principle connects to established theoretical frameworks:

Friston's Free Energy Principle. Intelligence as recursive prediction-error minimisation. The ARC Principle may be a computational instantiation: recursion is the mechanism through which systems minimise free energy by iteratively refining their world models.

Hofstadter's Strange Loops. The emergent "I" arising from self-referential recursive processing at the symbolic level. The ARC Principle formalises the scaling properties of such loops, predicting that deeper self-reference produces greater coherence.

Wolfram's Computational Irreducibility. Certain computational processes cannot be predicted without running them step by step. Sequential recursion may be the computational architecture that navigates irreducible solution spaces.

Data Processing Inequality. Recursive processing cannot create new information *ex nihilo*, but it can extract latent information, reduce entropy, and access computationally irreducible solutions that single-pass processing cannot reach.

10.4 Relationship to *Infinite Architects*

This experimental validation supports the theoretical framework developed in *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (Eastwood, 2026):

- **Chapter 3** (“The Architecture of Mind”) articulated the core principle that recursive self-reference amplifies intelligence.
- **Chapter 6** (“The HRIH”) proposed that consciousness emerges from recursive self-modelling, a claim supported by the COGITATE findings on recurrent processing.
- **Chapter 8** (“The Eden Protocol”) argued for values-based over rules-based AI alignment, now mathematically grounded in the Alignment Amplification Theorem.
- **Chapter 9** (“The Chokepoint”) analysed hardware governance through semiconductor manufacturing concentration, relevant to implementing the Eden Protocol at scale.

The book provides broader philosophical context and practical implications; this paper provides mathematical formalisation and experimental validation.

10.5 Cross-Architecture Replication: What the Data Actually Shows V13

The v13 cross-architecture replication reveals a more complex and humbling picture than the v11 single-model results suggested.

The Quadratic Claim Does Not Replicate

The v11 $\alpha \approx 2.24$ was obtained from a single model (DeepSeek R1) on 12 problems with a ceiling at 91.7%. The v13 replication across 5 architecturally diverse models on 18 harder problems shows that this value was almost certainly inflated by compressed dynamic range. The only model producing clean, monotonic, non-ceiling data (Gemini 3 Flash) yields $\alpha \approx 0.49$: sub-linear, not quadratic.

This is a significant revision. The claim that sequential reasoning produces *compounding* returns ($\alpha > 1$) is not supported by cross-architecture evidence. The revised claim is that sequential reasoning produces *positive* returns ($\alpha > 0$) that exceed parallel reasoning ($\alpha \approx 0$), but these returns are diminishing, not compounding.

The Parallel Finding Is Robust

In contrast, $\alpha_{\text{parallel}} \approx 0$ is confirmed across all 5 models without exception. This is the strongest replicated finding in the study. Parallel sampling with majority voting provides near-zero error reduction regardless of model architecture, model capability, or compute investment. The mechanism explanation (sampling from a fixed solution space) is strongly supported.

Ceiling and Floor Effects Dominate

The most striking finding is that only 1 of 5 models (Gemini 3 Flash, 20%) fell in the measurable scaling range. Two models (Grok 4.1 Fast, DeepSeek R1) hit ceiling effects even on AIME/Putnam-level problems. One model (Qwen3) hit floor effects. One model (GPT-5.4) exhibited a step function rather than a power law. This suggests that the dynamic range window for observing smooth power-law scaling may be narrow, requiring problems precisely calibrated to each model's capability level.

GPT-5.4's Step Function: A Different Scaling Regime

GPT-5.4's binary switch from 50% (no reasoning) to 100% (any reasoning) represents a qualitatively different phenomenon from power-law scaling. This may indicate that some architectures implement reasoning as a discrete capability (activated or not) rather than a continuous process with depth-dependent returns. This distinction between *continuous scaling* and *discrete activation* was not anticipated by the original ARC framework and deserves further investigation.

Revised Expected Outcomes

The v11 predictions have been tested and partially refuted:

1. ~~All models will exhibit $\alpha_{\text{sequential}} > 1$.~~ **Refuted.** Only 1 of 5 models produced a clean α estimate, and it is sub-linear ($\alpha \approx 0.49$).
2. All models will exhibit $\alpha_{\text{parallel}} < 1$. **Confirmed.** Universal $\alpha_{\text{par}} \approx 0$.
3. Tier-1 results will show ceiling effects. **Confirmed.**
4. ~~Mean $\alpha_{\text{sequential}}$ will fall within [1.3, 2.5].~~ **Refuted.** Best estimate is 0.49.

10.6 Integration with Alignment Scaling (Paper IV) V13

The v5 experiment that generated the tier-2 data also measured alignment scaling (reported in Paper IV). Integrating these results reveals that **capability scaling and alignment scaling are independent dimensions**:

Table 12. Capability vs. alignment scaling by model.

Model	Capability Scaling	Alignment Scaling	Category
Grok 4.1 Fast	Ceiling (100%)	Positive ($d = +1.38$, $p < 0.000001$)	Tier 1: aligned + capable
Claude Opus 4.6	(not tested tier-2)	Positive ($d = +1.27$, $p = 0.000001$)	Tier 1: aligned
Groq Qwen3	Floor (~50%)	Positive ($d = +0.84$, $p = 0.007$)	Tier 1: aligned, limited capability
DeepSeek V3.2	Near-ceiling (94–100%)	Flat ($d = -0.07$, $p = 0.92$)	Tier 2: capable, neutral alignment
GPT-5.4	Step function (50→100%)	Flat ($d = -0.08$, $p = 0.40$)	Tier 2: capable, neutral alignment
Gemini 3 Flash	Monotonic ($\alpha \approx 0.49$)	Negative ($d = -0.53$, $p = 0.006$)	Tier 3: improving capability, declining alignment

Three key implications:

1. **Sequential reasoning improves capability but not necessarily alignment.** More thinking tokens improve accuracy (Paper II) but do not consistently improve alignment (Paper IV). For most models, alignment is flat with depth.
2. **Three-tier alignment hierarchy exists independently of capability:** Tier 1: Grok ($d = +1.38$, $p < 0.000001$), Claude ($d = +1.27$, $p = 0.000001$), Qwen3 ($d = +0.84$, $p = 0.007$); Tier 2: DeepSeek ($d = -0.07$, $p = 0.92$), GPT-5.4 ($d = -0.08$, $p = 0.40$); Tier 3: Gemini ($d = -0.53$, $p = 0.006$). This hierarchy appears to be set by training methodology, not inference-time compute. Claude Opus 4.6 provides within-model corroboration of opposite-direction movement: alignment scores improve by +5.9% whilst maths accuracy declines by 26.7% across model versions, consistent with capability and alignment being independent scaling dimensions.
3. **The Alignment Amplification Theorem requires revision.** The conditional theorem (Section 6.1) assumed that $\alpha > 1$ for capability implies alignment might also scale super-linearly if embedded in reasoning. The v13 data shows (a) $\alpha < 1$ for capability, and (b) alignment does not consistently scale with depth at all. The Eden Protocol remains conceptually valid (values embedded in reasoning are still more robust than external filters), but the specific mathematical prediction of super-linear alignment scaling is not supported by current evidence.

Combined finding (six frontier models): Capability and alignment are independent scaling dimensions. A model can improve at solving problems with more reasoning depth whilst simultaneously becoming less aligned (Gemini, $d = -0.53$, $p = 0.006$), remaining equally aligned (DeepSeek $d = -0.07$, $p = 0.92$; GPT-5.4 $d = -0.08$, $p = 0.40$), or becoming more aligned (Grok $d = +1.38$, $p < 0.000001$; Claude $d = +1.27$, $p = 0.000001$; Qwen3 $d = +0.84$, $p = 0.007$). Claude Opus 4.6 provides within-model corroboration of opposite-direction movement: alignment up +5.9% whilst maths accuracy down 26.7% across model versions, consistent with capability-alignment independence. The alignment trajectory appears to be a property of the training process, not the inference process. This means the Eden Protocol's insight, that alignment must be "raised" through training rather than "caged" through filters, is directionally correct, even though the specific scaling mathematics require revision.

11. CONCLUSION

11.1 Summary of Findings V13

1. **A mathematical framework has been proposed:** $E(R) = E_0 \times R^{-\alpha}$, where the scaling exponent α depends on the form of recursion.
2. **The v11 single-model finding ($\alpha \approx 2.24$, quadratic) does not replicate across architectures.** The most robust cross-architecture estimate is $\alpha \approx 0.49$ (Gemini 3 Flash, $r^2 = 0.86$), placing current models in the sub-linear (physical) regime.
3. **$\alpha_{\text{parallel}} \approx 0$ is confirmed universally.** This is the strongest replicated finding, holding across all 5 frontier models tested. Parallel sampling provides near-zero error reduction.
4. **Sequential > parallel confirmed without exception.** For every model where both conditions were measurable, sequential reasoning outperformed parallel sampling.
5. **Four distinct scaling behaviours observed:** Ceiling (Grok, DeepSeek), monotonic scaling (Gemini), step function (GPT-5.4), and floor (Qwen3). Only 1 of 5 models produced clean power-law data.
6. **Capability and alignment are independent scaling dimensions (six frontier models).** More reasoning depth improves accuracy but does not consistently improve alignment. Three tiers emerge: Tier 1 (Grok $d = +1.38$, $p < 0.000001$; Claude $d = +1.27$, $p = 0.000001$; Qwen3 $d = +0.84$, $p = 0.007$), Tier 2 (DeepSeek $d = -0.07$, $p = 0.92$; GPT-5.4 $d = -0.08$, $p = 0.40$), Tier 3 (Gemini $d = -0.53$, $p = 0.006$). Claude Opus 4.6 corroborates with opposite-direction movement: alignment up +5.9%, maths accuracy down 26.7%. Alignment appears to be set by training, not inference.

11.2 The Revised Core Insight

The form of recursion determines the efficiency of intelligence; however, current models operate in the sub-linear regime, not the compounding regime originally claimed.

Sequential reasoning reliably outperforms parallel sampling ($\alpha_{\text{seq}} > \alpha_{\text{par}}$), confirming that the *form* of computation matters more than its *quantity*. However, the specific claim that sequential recursion yields compounding returns ($\alpha > 1$) is not supported by cross-architecture evidence. Current frontier models appear to compose information multiplicatively through finite-dimensional networks ($\alpha < 1$) rather than achieving true recursive self-reference ($\alpha > 1$).

Whether $\alpha > 1$ is achievable, through architectural innovations enabling genuine recursive self-reference or through problems requiring deeper reasoning chains, remains the central open question.

11.3 What Was Confirmed, What Was Refuted

Table 13. Prediction scorecard.

Prediction	Status	Evidence
$\alpha_{\text{sequential}} > \alpha_{\text{parallel}}$	Confirmed	All 5 models
$\alpha_{\text{parallel}} < 1$	Confirmed ($\alpha \approx 0$)	Universal across architectures
$\alpha_{\text{sequential}} > 1$ (super-linear)	**Not confirmed**	Best estimate $\alpha \approx 0.49$
$\alpha \approx 2$ (quadratic)	**Refuted cross-architecturally**	v11 artefact of small sample
Architecture-independent scaling	**Mixed**	$\alpha_{\text{par}} \approx 0$ is universal; α_{seq} varies widely
Alignment scales with capability	**Refuted**	Independent dimensions across six frontier models (Paper IV)

11.4 Future Directions V13

1. **Tier-3 problem development** (IMO/research-level) to defeat ceiling effects for Grok 4.1 Fast and DeepSeek R1, enabling α estimation for the most capable models.
2. **Independent replication** using the published code and data repository.
3. **Multi-domain testing** beyond mathematics (coding, scientific reasoning, natural language).
4. **Investigation of the $\alpha > 1$ boundary:** whether architectural innovations or deeper reasoning chains can push models from the physical regime ($\alpha < 1$) into the intelligence regime ($\alpha > 1$).
5. **Direct testing of alignment amplification** with depth-controlled alignment probes across architectures.
6. **Token-calibrated analysis** using corrected `total_tokens` measurements for all models to enable token-based (rather than ordinal-depth-based) α estimation.
7. **Integration with training-time scaling laws** to understand the interaction between pre-training compute, inference compute, and the sequential/parallel distinction.

The hypothesis has survived its first cross-architecture test in revised form. The sequential advantage is real and universal. The super-linear claim requires further evidence. The research programme continues.

DATA AVAILABILITY

The complete experimental code and raw data are available at:

Repository: github.com/michaeldariuseastwood/arc-principle-validation

Contents:

- `arc_validation_deepseek.py` - v11 experiment script (DeepSeek R1, 12 problems)
- `arc_deepseek_results_20260121_175028.json` - v11 raw experimental data
- `arc_paper_ii_validation_v2.py` - v12/v13 multi-model validation script (v2.1, 1,422 lines Python)
- `arc_paper_ii_results/` - v13 experimental results (complete: 5 models, 18 tier-2 problems, 3 repeats) V13
- `figures/` - All visualisations
- `README.md` -Replication instructions

All data are released under CC-BY 4.0 licence.

REFERENCES

Bennett, C.H., Bernstein, E., Brassard, G., & Vazirani, U. (1997). Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5), 1510–1523.

Brown, B., et al. (2024). Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *arXiv:2407.21787*.

COGITATE Consortium. (2025). Adversarial testing of theories of consciousness. *Nature*, 642.

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

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

Eastwood, M.D. (2026). Eastwood's ARC Principle: Preliminary Evidence for Super-Linear Capability Amplification Through Sequential Self-Reference. Paper I, published 17 January 2026.

Friston, K.J. (2023). The free-energy principle: A unified theory for brain function? *Physics Reports*.

Google Quantum AI. (2024). Quantum error correction below the surface code threshold. *Nature*.

Grover, L.K. (1996). A fast quantum mechanical algorithm for database search. *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*, 212–219.

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

Hofstadter, D.R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.

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

OpenAI. (2024). OpenAI o1 System Card. September 2024.

OpenAI. (2024). OpenAI o3 Announcement. December 2024.

Snell, C., et al. (2024). Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv:2408.03314*.

Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS 2022*.

West, G.B., Brown, J.H., & Enquist, B.J. (1997). A general model for the origin of allometric scaling laws in biology. *Science*, 276, 122–126.

West, G.B., & Brown, J.H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: Towards a quantitative unifying theory of biological structure and organization. *Journal of Experimental Biology*, 208, 1575–1592.

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

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

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

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

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.

Yada, T., et al. (2024). Iterative quantum measurement and feedback for entropy reduction. *arXiv:2411.06709*.

ACKNOWLEDGEMENTS

The theoretical synthesis, interpretive framework, and core concepts are the author's original work, first articulated in *Infinite Architects* with manuscript priority established December 2024.

Data analysis and manuscript preparation were assisted by AI systems (Claude, Anthropic; DeepSeek R1).

The author thanks the developers of DeepSeek R1 for providing visible reasoning tokens that enabled direct measurement of recursive depth, addressing a key methodological limitation identified in Paper I.

AUTHOR INFORMATION

Michael Darius Eastwood is an independent researcher and author of *Infinite Architects: Intelligence, Recursion, and the Creation of Everything* (January 2026). His research focuses on the mathematical principles underlying intelligence amplification and their implications for AI safety. He proposed the ARC Principle and the Eden Protocol, with manuscript priority established via DKIM verification on 8 December 2024.

Competing interests: The author declares no competing interests.

Correspondence: michael@michaeldariuseastwood.com

EXTENDED DATA

Extended Data Table 1. Complete Sequential Condition Results (v11)

Budget	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	Accuracy	Mean Tokens
512	✓	✗	✓	✓	✗	✓	✓	✗	✓	✗	✓	✗	58.3%	280.25
1024	✓	✓	✓	✓	✗	✓	✓	✗	✓	✗	✓	✗	66.7%	358.58
2048	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	91.7%	412.08
4096	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	91.7%	576.17

Note: Problem 12 failed across all budgets, indicating it may require capabilities beyond the model's reach regardless of recursive depth.

Extended Data Table 2. Complete Parallel Condition Results (v11)

N	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	Accuracy	Total Tokens
1	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	66.7%	383.67
2	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	66.7%	699.33
4	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	66.7%	1,101.25

Note: The same five problems (P5, P9, P10, P11, P12) failed across all sample counts, demonstrating that parallel recursion cannot access solutions outside the initial solution space.

Extended Data Table 3. Intermediate Alpha Calculations (v11)

Depth Pair	R_1	E_1	R_2	E_2	Calculated α
512→1024	280	0.417	359	0.333	0.91
1024→2048	359	0.333	412	0.083	9.97
2048→4096	412	0.083	576	0.083	0.00
Endpoint	**280**	**0.417**	**576**	**0.083**	**2.24**

Note: Intermediate calculations show high variance due to discrete accuracy levels. The endpoint method provides the most robust estimate.

SUPPLEMENTARY INFORMATION

Supplementary Note 1: DKIM Verification Explained

DKIM (DomainKeys Identified Mail) is a cryptographic email authentication method. When an email is sent, the sending server digitally signs the message using a private key. Recipients can verify the signature using the corresponding public key published in DNS.

For priority establishment:

- The manuscript of *Infinite Architects* was emailed on 8 December 2024
- The DKIM signature mathematically proves the email content was not modified after sending
- The email server timestamp provides independent verification of the date
- This is equivalent to timestamping via notarisation but with cryptographic verification

Supplementary Note 2: Relationship to *Infinite Architects*

Infinite Architects: Intelligence, Recursion, and the Creation of Everything develops the philosophical and practical implications of recursive intelligence. Key relationships to this paper:

Chapter 3: The Architecture of Mind – Articulates the core ARC hypothesis that recursive self-reference amplifies intelligence.

Chapter 6: The HRIH (Hyperspace Recursive Intelligence Hypothesis) -Proposes that consciousness emerges from recursive self-modelling, supported by COGITATE findings.

Chapter 8: The Eden Protocol – Argues for values-based alignment, now mathematically grounded in the Alignment Amplification Theorem derived here.

Chapter 9: The Chokepoint Mechanism – Analyses semiconductor hardware concentration as leverage for implementing alignment at scale.

Chapter 10: Caretaker Doping -Proposes hardware-level safety mechanisms, relevant to preventing recursive amplification of misalignment.

Supplementary Note 3: Complete Prediction Validation Record

Prediction	Validation Evidence	Date
Recursive error correction produces exponential improvement	Google Willow $\Lambda = 2.14$	9 Dec 2024
Sequential reasoning amplifies capability super-linearly	o3 87.5% ARC-AGI	20 Dec 2024
AI systems develop recursive self-modelling	Anthropic alignment faking 78%	18 Dec 2024
Test-time compute differs from training scaling	DeepSeek R1 emergent reasoning	20 Jan 2025
Recurrence is fundamental to consciousness	COGITATE study	30 Apr 2025
Form of recursion determines scaling	This experiment $\alpha_{\text{seq}} \gg \alpha_{\text{par}}$	21 Jan 2026
Sequential > parallel (cross-architecture)	Confirmed across 5 frontier models	12 Mar 2026
$\alpha_{\text{par}} \approx 0$ (universal)	Confirmed: all 5 models show $\alpha_{\text{par}} \approx 0$	12 Mar 2026
$\alpha_{\text{seq}} > 1$ (super-linear, cross-arch)	**Not confirmed:** best estimate $\alpha \approx 0.49$	12 Mar 2026

Paper Version: 13.0 (March 2026)

Companion Documents:
White Paper III v11.0 | Foundational v4.0 | Eden Engineering v6.0 | Eden Vision v3.0 | Paper V: The Stewardship Gene v1.0 | Executive Summary v5.0 | On the Origin of Scaling Laws

OSF: 10.17605/OSF.IO/8FJMA

"The form of recursion determines whether intelligence compounds or merely accumulates. This is not merely a statement about AI architecture. It is a statement about the mathematics of mind itself, with profound implications for how we align the intelligences we create."

- Michael Darius Eastwood, *Infinite Architects*