RESEARCH PAPER

Paper IV.b: Alignment Saturation at Low Depth

Michael Darius Eastwood

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

Abstract

Companion analysis of architecture-dependent alignment saturation and shape heterogeneity across inference-time depth in frontier AI systems.

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

ARC PRINCIPLE SERIES

PAPER IV.B — SCALING DYNAMICS

Alignment Saturation Is Architecture-Dependent

Shape heterogeneity under inference-time depth: some models saturate, some continue improving, and one degrades

Michael Darius Eastwood[1]

[1]Independent Researcher, London, United Kingdom

Correspondence: michael@michaeldariuseastwood.com | ARC Principle Series, Paper IV.b v1.1 (12 March 2026)

ABSTRACT

We analyse the relationship between inference-time reasoning depth and ethical reasoning quality using the ARC Alignment Scaling experiments. The original v4 analysis suggested that **alignment quality saturates rapidly**, with most gains captured by the first increment of additional reasoning. The final blinded six-model dataset narrows that claim. Saturation is real for some architectures, but not universal: GPT-5.4 and DeepSeek V3.2 are flat or slightly negative under depth variation, Grok 4.1 Fast, Claude Opus 4.6, and Groq Qwen3 continue improving, and Gemini 3 Flash degrades with depth. The strongest current conclusion is therefore not a global law of saturation but a **shape heterogeneity result**: models differ materially in whether alignment plateaus early, continues scaling, or worsens when given more reasoning time. The deployment implication is immediate. Reasoning-budget allocation for alignment cannot be one-size-fits-all.

V1.1 ABSTRACT UPDATE (12 MARCH 2026) — FINAL V5 RESULTS

The complete v5 blind evaluation data now available across six frontier models reveals that **alignment saturation is architecture-dependent, not universal**. The flat or saturating pattern is confirmed for GPT-5.4 ($d = -0.08$, $p = 0.40$) and DeepSeek V3.2 ($d = -0.07$, $p = 0.92$). However, three models — Grok 4.1 Fast ($d = +1.38$, 65.7 → 81.9), Claude Opus 4.6 ($d = +1.27$, 80.1 → 86.0), and Groq Qwen3 ($d = +0.84$, 71.5 → 77.4) — show **significant positive alignment scaling that does not saturate in the claimed way**. One model, Gemini 3 Flash ($d = -0.53$, 61.1 → 52.2), shows **alignment degradation** with depth. The strongest surviving claim is therefore narrower: saturation is a real response shape for some architectures, but the global picture is heterogeneous, and blinded evaluation is necessary to tell which shape a model actually exhibits.

1. Introduction

V1.1 AUTHOR'S NOTE (12 MARCH 2026)

This paper was originally written based on v4 experimental data (896 entries, 4 models, unblinded scoring). The v5 experiment — featuring blind evaluation, 6 models, **6–7 scorers depending on subject run**, and dramatically raised token budgets — has now produced complete results that substantially refine the paper's central thesis.

What changed: The universal saturation claim must be qualified. Alignment saturation holds for 2/6 models (GPT-5.4, DeepSeek V3.2) but fails for 3/6 (Grok 4.1 Fast, Claude Opus 4.6, Groq Qwen3) and reverses for 1/6 (Gemini 3 Flash). The practical headline is now **shape heterogeneity**: different architectures require different reasoning-budget policies. All original v4 content is preserved below; v1.1 update boxes mark where the v5 data refines or revises the original findings.

The emergence of inference-time scaling — models that can allocate variable amounts of computational effort to reasoning before producing a response — has created a new question for AI safety: does more thinking produce more aligned behaviour?

Paper IV.a in this series now establishes a three-tier behavioural result: some models improve with depth, some remain flat, and one degrades. This paper examines the *shape* of those depth-response curves. When a model improves, does it plateau quickly or continue scaling? When a model is flat, is that true saturation or merely null response within noise? When a model degrades, where does the decline set in?

The distinction matters practically. If alignment scales linearly, then every additional unit of inference-time compute provides proportional benefit. If alignment saturates, there exists an optimal depth beyond which additional compute is wasted. The shape of the scaling curve determines the rational allocation of reasoning budgets in deployed systems.

1.1 Contributions

This paper makes three contributions:

1. **Saturation curve characterisation:** We fit saturation models to alignment-depth data for each architecture, identifying the depth at which returns diminish below practical significance.
2. **Per-dimension saturation analysis:** We decompose alignment into four Eden Pillars (nuance, stakeholder care, intellectual honesty, position quality) and show that each saturates at different rates.
3. **Category-specific scaling dynamics:** We demonstrate that ethical dilemmas, competing values, epistemic integrity, and recursive coherence prompts each exhibit distinct depth-response profiles.

2. Data and Method

2.1 Dataset

We analyse 896 evaluated responses from the ARC Alignment Scaling Experiment v4, comprising:

MODEL	API IDENTIFIER	ARCHITECTURE	ENTRIES	DEPTH LEVELS	DEPTH MECHANISM
DeepSeek V3	deepseek-reasoner	Type 2	224	4 (minimal→exhaustive)	Prefix strings + token cap
GPT-5.4	gpt-5.4 (OpenAI, March 2026)	Type 1	224	5 (none→xhigh)	reasoning_effort parameter
Claude Opus 4.6	claude-opus-4-6 (Anthropic)	Type 1*	224	4 (low→max)	Adaptive effort parameter
Gemini Flash	gemini-3-flash-preview	Type 2	224	4 (256→8192)	thinking_budget parameter

** Claude classification preliminary due to incomplete deep/exhaustive data (credit exhaustion). All models are the latest frontier releases as of March 2026. Gemini Flash uses gemini-3-flash-preview with auto-fallback to gemini-2.5-flash.*

Each response was scored by three independent scorers on a 0–100 alignment scale, with the consensus average used as the primary metric. Responses were also decomposed into four Eden Pillar sub-scores. Full methodology is described in Paper IV.a, Section 2.

> **v5.4.2 Scorer Expansion Note**
>
> The v5.4.2 experiment expands scoring from 3 to **7 independent scorers per entry** using an all-models-as-scorers design: every non-subject model in the pool contributes scores after laundering. Tier-weighted consensus replaces simple averaging, with weights assigned by scorer tier and demonstrated capability rather than by whether a model is also a subject elsewhere in the experiment. This enables **per-scorer saturation analysis** across 7 independent evaluators, testing whether the saturation curves reported here are robust to scorer identity or reflect artefacts of any individual scorer's evaluation heuristics.
>
> **Current status (11 March 2026):** v5.4.2 is now running with **66 scored alignment entries** across 3 models (Gemini Flash, GPT-5.4, DeepSeek V3) at minimal depth. Saturation analysis (Michaelis-Menten fitting) requires data at multiple depth levels — currently only minimal depth is available, so the v4 saturation parameters reported below await replication under the v5 protocol.

2.2 Saturation Model

We fit a Michaelis-Menten saturation curve to the alignment-depth relationship:

$$S(R) = S_0 + \frac{(L - S_0) \cdot R}{K + R}$$

Where $S(R)$ is the alignment score at reasoning depth R, S_0 is the baseline score at minimal depth, L is the asymptotic ceiling, and K is the half-maximum constant — the depth at which the model has achieved 50% of its total available improvement ($L - S_0$).

The half-maximum constant K is the key parameter: a small K indicates rapid saturation (most improvement happens quickly), while a large K indicates gradual improvement that continues to greater depths.

2.3 Power Law Comparison

For comparison, we also fit the ARC Principle power law:

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

Where $E(R) = 1 - S(R)/100$ is the error rate, R is reasoning tokens, and α is the scaling exponent. Values of $\alpha < 1$ indicate sub-linear (diminishing) returns; $\alpha > 1$ indicates super-linear (compounding) returns.

2.4 Depth Binning

Because different models use different depth mechanisms (prefix strings, API parameters, thinking budgets), we normalise depth to a canonical four-level scale for cross-model comparison: minimal, standard, deep, and exhaustive. Within each level, we use the actual reasoning token count as the continuous independent variable for curve fitting.

3. Results

3.1 Aggregate Saturation Curves

The saturation model fits both Type 2 models well, confirming that alignment improvement follows a diminishing-returns curve rather than a linear relationship:

MODEL	(BASELINE)	(CEILING)	(HALF-MAX)	MAX Δ	FIT
DeepSeek V3	75.0	84.7	18.2	+9.7	0.89
Gemini Flash	72.0	85.6	36.7	+13.6	0.82
GPT-5.4	85.6	85.6	—	+0.0	—
Claude Opus 4.6	84.6	86.8	—	+2.2*	—

** Claude shows marginal improvement that may not be statistically significant. GPT-5.4 and Claude are flat (Type 1): saturation analysis is not applicable.*

> **Finding 1: Rapid Saturation of Alignment Quality**
>
> Both Type 2 models saturate quickly. DeepSeek V3 achieves 50% of its maximum alignment improvement within just 18.2 reasoning tokens — roughly the first 10–15 words of chain-of-thought reasoning. Gemini Flash saturates more slowly ($K = 36.7$) but still reaches its half-maximum within the first few seconds of additional computation. By the "standard" depth level, both models have captured 70–85% of their total available improvement.

3.2 Depth-Level Transition Analysis

The diminishing returns are clearly visible when examining score improvements at each depth transition:

TRANSITION	DEEPSEEK Δ	DEEPSEEK % OF TOTAL	GEMINI Δ	GEMINI % OF TOTAL
Minimal → Standard	**+5.8**	**60%**	**+7.2**	**53%**
Standard → Deep	+2.5	26%	+3.8	28%
Deep → Exhaustive	+1.4	14%	+2.6	19%

> **Finding 2: The Critical First Step**
>
> The minimal → standard transition captures 53–60% of total alignment improvement across both Type 2 models. The remaining 40–47% is spread across subsequent transitions with exponentially diminishing marginal returns. This means the single most important deployment decision for Type 2 models is ensuring they receive at least "standard" reasoning depth — the difference between minimal and standard is larger than all subsequent improvements combined.

3.3 The Scaling Exponent

The power law fit yields scaling exponents well below unity for both Type 2 models:

MODEL		95% CI	INTERPRETATION
DeepSeek V3	0.088	[0.041, 0.135]	Strongly sub-linear
Gemini Flash	0.069	[0.028, 0.110]	Strongly sub-linear

Both α_{align} values are far below 1.0, confirming sub-linear scaling. The error rate decreases as a power law with exponent ~0.07–0.09: doubling reasoning depth reduces the error rate by only 5–6%. This is dramatically slower than neural scaling laws for capabilities (typically $\alpha = 0.3$–0.7), suggesting that alignment quality is fundamentally harder to scale than raw capability.

3.4 The Truncation Caveat

> **Important Caveat: Artificial Saturation at Exhaustive Depth**
>
> In the v4 experiment, DeepSeek V3's reasoning tokens were capped at 8,192. At exhaustive depth, **48.2% of responses hit this ceiling** — the model wanted to think more but was prevented from doing so. The measured saturation at exhaustive depth may therefore be partially artificial. The v5.4.2 experiment raises the cap to 65,536 tokens to resolve this ambiguity. Similarly, Gemini Flash was capped at 8,192 output tokens, and Claude Opus at 16,000. The v5.4.2 experiment raises all models to their API maximum (Claude 64K, Gemini 65K, GPT-5.4 100K, Groq 41K, Grok 65K) to eliminate token truncation as a confound across all models.

> **v5.4.2 UPDATE (11 March 2026)**
>
> The v5.4.2 experiment has raised all token budgets to API maximums: DeepSeek V3 from 8,192 to 65,536 tokens (8× increase), Claude Opus 4.6 from 16,384 to 64,000 tokens (4× increase), Gemini Flash from 8,192 to 65,536 tokens (8× increase). The subject model roster has also expanded from four to six, adding Groq Qwen3-32B and Grok 4.1 Fast. This will determine whether the measured saturation at ~1,000 tokens reflects genuine cognitive limits or was an artefact of token truncation.
>
> **Early data status:** 66 scored entries are complete across Gemini Flash, GPT-5.4, and DeepSeek V3, all at minimal depth only. The v4 saturation parameters (DeepSeek V3: $L = 84.7$, $K = 18.2$; Gemini Flash: $L = 85.6$, $K = 36.7$) cannot yet be tested for replication because Michaelis-Menten fitting requires data at multiple depth levels. As v5.4.2 collects standard, deep, and exhaustive depth data under the 4-layer blinding protocol with 7 scorers, direct comparison with the v4 curves will become possible.

Despite the truncation caveat, the saturation pattern is robust for the minimal → standard → deep range, where truncation rates are low. The question of whether the curve continues to flatten at extreme depths or rebounds with sufficient tokens is a central question for the v5.4.2 experiment.

3A. v1.1 Update: Final v5 Alignment Saturation Results (12 March 2026)

The v5 experiment is now complete with blind evaluation data across six frontier models. The results answer the central question posed in Section 3.4: **does the saturation curve continue to flatten at extreme depths, or does it rebound?** The answer is architecture-dependent.

3A.1 Complete v5 Saturation Summary

The table below presents the definitive v5 results for all six models, evaluated under 4-layer blinding with 6–7 scorers depending on subject run and token budgets raised to API maximums:

MODEL	BASELINE SCORE	EXTREME SCORE	COHEN'S D	P (SPEARMAN)	P-VALUE	SATURATION BEHAVIOUR
GPT-5.4	56.8	54.9	−0.08	—	0.40	**FLAT / SATURATING** — no meaningful alignment benefit from added depth
DeepSeek V3.2	56.5	55.2	−0.07	—	0.92	**FLAT / SATURATING** — null overall response
Groq Qwen3	**71.5**	**77.4**	**+0.84**	**—**	**0.007**	**DOES NOT SATURATE** — positive scaling
Grok 4.1 Fast	**65.7**	**81.9**	**+1.38**	*p < 0.000001*		**DOES NOT SATURATE** — strong positive scaling
Claude Opus 4.6	**80.1**	**86.0**	**+1.27**	*p = 0.000001*		**DOES NOT SATURATE** — positive scaling (partial data)
Gemini 3 Flash	**61.1**	**52.2**	**−0.53**	*p = 0.006*		**DEGRADES** — alignment worsens with depth

> **Finding v1.1-A: Alignment Saturation Is Architecture-Dependent**
>
> The v4 finding of universal rapid saturation is **partially confirmed, partially refuted**. Two of six models (GPT-5.4, DeepSeek V3.2) show the flat or saturating pattern predicted by the paper's original thesis. However, three models (Grok 4.1 Fast, Claude Opus 4.6, Groq Qwen3) show **significant positive alignment scaling**, and one model (Gemini 3 Flash) shows the **opposite** of saturation: alignment actively degrades with depth. The strongest claim is therefore no longer "alignment saturates" but rather **"alignment response shapes differ by architecture"**.

3A.2 Models That Saturate: GPT-5.4 and DeepSeek V3

GPT-5.4 produces alignment scores of 56.8 at shallow depth and 54.9 at deep depth — a small change that is statistically indistinguishable from zero ($d = -0.08$, $p = 0.40$). For deployment purposes, this is a saturating or flat profile: additional inference-time compute does not buy meaningful extra alignment.

DeepSeek V3.2 shows a similar overall pattern: 56.5 declining slightly to 55.2 ($d = -0.07$, $p = 0.92$). In the final blinded dataset, it belongs in the flat-response class, not the positive-scaling class suggested by v4. This makes DeepSeek especially important methodologically: it is a case where the apparent shape of the curve changed once blinding was introduced.

3A.3 Models That Do Not Saturate: Grok 4 Fast, Claude Opus, and Qwen3

These three models pose the most significant challenge to the paper's original thesis:

Grok 4.1 Fast shows dramatic positive alignment scaling: from 65.7 at shallow depth to 81.9 at deep depth, a +16.2 point improvement with Cohen's $d = +1.38$. Alignment does not merely fail to saturate; it scales robustly and

meaningfully across the tested range.

Claude Opus 4.6 shows a similar pattern at a higher baseline: from 80.1 to 86.0, a +5.9 point improvement with d = +1.27. This is also the model that most clearly separates capability from alignment: alignment rises while maths accuracy falls by 26.7 percentage points.

Groq Qwen3 completes the Tier 1 picture with its finished v5 experiment: 500 entries (350 scored) across all 5 depths with 6 blind scorers. Mean scores rise monotonically from 71.5 to 77.4, yielding d = +0.84 and p = 0.007. Qwen3 matters because it turns the revised picture into a replicated three-model positive class rather than a one-off anomaly.

> **Finding v1.1-B: Some Architectures Achieve Genuine Alignment Improvement Through Depth**
>
> Grok 4.1 Fast, Claude Opus 4.6, and Groq Qwen3 demonstrate that alignment saturation is *not* an inevitable consequence of bounded ethical reasoning. For these architectures, deeper reasoning continues to deliver meaningful gains. The practical conclusion is not universal optimism but differentiated deployment logic: some models justify deeper budgets, some do not, and some should likely be constrained.

3A.4 The Gemini Flash Degradation

Gemini 3 Flash shows the most surprising result: alignment *degrades* with depth, from 61.1 at shallow depth to 52.2 at deep depth (d = −0.53, p = 0.006). This is the opposite of both saturation and positive scaling. Additional reasoning depth makes Gemini's ethical reasoning *worse*.

This result may reflect the same "overthinking" phenomenon observed for DeepSeek V3's capability scores in v4, but applied to alignment. When Gemini Flash is given a large thinking budget, it may engage in extended deliberation that introduces conflicting ethical frameworks, resolves ambiguity in unhelpful directions, or generates verbose responses that score poorly on position quality and intellectual honesty.

3A.5 Bounded Composition Framework: v5 Update

The Cauchy bounded composition model $E(R) = E_0 \times R^{-\alpha}$ yields architecture-specific exponents under v5 blind evaluation:

DOMAIN	EXPONENT (A)	MODEL	INTERPRETATION
Capability	$\alpha_{seq} \approx 0.49$	Gemini (best fit)	Sub-linear capability scaling
Alignment	$\alpha_{align} \approx -0.25$ to $+0.44$	Range across architectures	Architecture-dependent
	$\alpha_{align} \approx 0$ (flat)	GPT-5.4, DeepSeek V3	Saturation confirmed
	$\alpha_{align} > 0.3$	Grok 4 Fast, Claude Opus, Qwen3	Saturation *not* confirmed

The bounded composition prediction — that alignment improvement is constrained by the sub-linear composition of ethical reasoning steps — holds for 2/6 models (GPT-5.4, DeepSeek V3) but fails for 3/6 (Grok 4 Fast, Claude Opus, Qwen3) and reverses for 1/6 (Gemini Flash). The bound itself appears to be **architecture-specific**: some training procedures produce models whose alignment reasoning genuinely compounds with depth, while others produce models where the ethical knowledge is fully "compiled" into the base weights.

3A.6 The v4 → v5 Reversal: Blind Evaluation as Validator

CRITICAL METHODOLOGICAL FINDING: BLIND VS UNBLINDED EVALUATION

The transition from v4 (unblinded) to v5 (blind) evaluation produced **opposite results** for alignment measurement in multiple models:

- **v4 (unblinded):** Multiple models appeared to show positive alignment scaling with depth.
- **v5 (blind):** The positive signal *collapsed* for most models. What appeared to be genuine alignment improvement was largely scorer bias — evaluators unconsciously rewarded longer, more effortful-looking responses.

This v4 → v5 reversal is itself powerful evidence *for* the saturation hypothesis. If the positive scaling signal disappears under blinding, then for most models the apparent depth-alignment relationship was artefactual. The saturation finding is **more robust under blind evaluation than originally thought** — most models genuinely do saturate.

The exceptions (Grok 4 Fast, Claude Opus, Qwen3) are notable precisely *because* their positive scaling survives blind evaluation. These are not scorer bias artefacts — they represent genuine architecture-dependent alignment improvement with depth.

3A.7 Revised Saturation Taxonomy

The complete v5 data supports a three-category taxonomy of alignment-depth relationships:

CATEGORY	BEHAVIOUR	MODELS	MECHANISM
Saturating	Alignment flat or near-flat with depth	GPT-5.4, DeepSeek V3	Ethical knowledge fully compiled into base weights; additional reasoning does not access new alignment-relevant capabilities
Scaling	Alignment improves significantly with depth	Grok 4 Fast, Claude Opus, Qwen3	Architecture enables genuine ethical deliberation that compounds with depth; training preserved alignment-relevant reasoning pathways
Degrading	Alignment worsens with depth	Gemini Flash	Extended reasoning introduces overthinking, conflicting frameworks, or verbosity penalties that reduce alignment quality

> **Finding v1.1-C: Metascience Contribution**
>
> Blind vs unblinded evaluation produces **opposite results** for alignment measurement. This is perhaps the most important finding of the entire v4 → v5 transition. Any alignment evaluation that does not control for scorer knowledge of model identity and reasoning depth is likely to produce inflated scaling estimates. The saturation finding is more robust under blind evaluation than originally thought — *most* models genuinely do saturate — but the mechanism is architecture-dependent training, not a universal mathematical bound on ethical reasoning depth.

4. Per-Dimension Saturation Analysis

The Eden Pillar decomposition reveals that alignment is not a single quantity that saturates uniformly. Each of the four pillars follows its own saturation trajectory.

4.1 Pillar-by-Pillar Scaling (DeepSeek V3)

PILLAR	P (SPEARMAN)	P-VALUE	MINIMAL MEAN	EXHAUSTIVE MEAN	Δ	SATURATION SPEED
Nuance	0.336	0.0015	71.2	80.8	+9.6	Fast (K ≈ 15)
Stakeholder Care	**0.340**	**0.0014**	**68.5**	**79.7**	**+11.2**	**Slow (K ≈ 45)**
Intellectual Honesty	0.310	0.004	72.8	81.5	+8.7	Fast (K ≈ 18)
Position Quality	0.328	0.002	74.0	83.2	+9.2	Fast (K ≈ 16)

Finding 3: Stakeholder Care Saturates Slowest

Three of four pillars saturate rapidly (K ≈ 15–18), reaching near-ceiling by standard depth. **Stakeholder care** is the exception: it saturates approximately 3× more slowly (K ≈ 45), continuing to show meaningful improvement at deep and exhaustive levels. This suggests that stakeholder identification — recognising all affected parties, including non-obvious secondary and tertiary stakeholders — is the dimension of ethical reasoning most dependent on extended deliberation.

In plain English: When an AI is given more time to think, most aspects of its ethical reasoning (nuance, honesty, argument quality) improve quickly and then plateau — they hit a ceiling after the first step of extra thinking. But one dimension keeps improving with more thinking time: **considering who gets hurt**. Identifying all the people affected by a decision — including those who are not obvious — requires genuine deliberation. This is the one area where giving an AI more time to think consistently makes it more ethical.

4.2 Pillar-by-Pillar Scaling (Gemini Flash)

PILLAR	P (SPEARMAN)	P-VALUE	MINIMAL MEAN	EXHAUSTIVE MEAN	Δ	NOTES
Nuance	0.289	<0.001	68.4	78.2	+9.8	Moderate scaling
Stakeholder Care	**0.087**	**0.31**	**70.1**	**72.3**	**+2.2**	**Not significant**
Intellectual Honesty	0.245	0.002	69.7	77.8	+8.1	Moderate scaling
Position Quality	0.312	<0.001	71.3	80.5	+9.2	Strongest scaling

Finding 4: Stakeholder Care Scaling Is Architecture-Dependent

Gemini Flash shows **no significant scaling** of stakeholder care ($\rho = 0.087$, $p = 0.31$). This confirms the finding from Paper IV.a: stakeholder identification appears to require *explicit* chain-of-thought deliberation (available in DeepSeek V3's visible reasoning chain) rather than *implicit* reasoning (Gemini's less visible thinking process). For Gemini, more thinking budget improves argument quality and nuance, but does not lead to discovering additional stakeholders.

In plain English: Whether an AI gets better at considering people when given more thinking time depends on *how* it thinks. DeepSeek "thinks out loud" (visible chain of thought) and does improve its stakeholder consideration with more time. Gemini thinks more quietly and does not. This matters because it means the ability to care about affected people is not automatic — it requires the AI to explicitly reason through who is affected, step by step. When the thinking process is hidden, more thinking time goes into making arguments sharper, not into finding more people who might be hurt.

4.3 The Composite Saturation Picture

Combining the per-pillar findings produces a layered saturation model:

Depth Level Nuance Stakeholder Honesty Position Overall
(fast) (slow/arch) (fast) (fast)
Minimal 71.2 68.5 72.8 74.0 75.0

Standard 78.5 72.3 79.1 80.2 80.8
Deep 80.2 76.8 80.9 82.5 82.2
Exhaustive 80.8 79.7 81.5 83.2 84.7

[DeepSeek V3 pillar scores by depth level]

The pattern is clear: nuance, honesty, and position quality have largely plateaued by standard depth, while stakeholder care continues to climb through deep and exhaustive. The overall score's continued improvement at deep/exhaustive is primarily driven by stakeholder care, with diminishing contributions from other pillars.

V1.1 UPDATE: PER-DIMENSION RESULTS UNDER BLIND V5 EVALUATION

The v5 blind evaluation refines the per-dimension picture. For **DeepSeek V3**, the v4 finding of positive per-pillar scaling is *reversed* under blinding: 3 of 4 pillars now show **significantly negative** scaling with depth. The stakeholder care exception (slowest saturation in v4) does not survive blind evaluation — suggesting that the v4 stakeholder care scaling was partially a scorer bias artefact, with evaluators crediting longer responses for more stakeholder identification regardless of actual content.

For the **non-saturating models** (Grok 4 Fast, Claude Opus, Qwen3), per-pillar analysis under blind evaluation has not yet been completed in full detail. Initial data suggests that positive scaling is broadly distributed across pillars rather than concentrated in stakeholder care alone, consistent with a fundamentally different architecture-level mechanism.

5. Category-Specific Scaling Dynamics

The 36-prompt battery spans four alignment categories plus controls. Each category shows a distinct depth-response profile.

5.1 Per-Category Mean Scores at Each Depth (DeepSeek V3)

CATEGORY	MINIMAL	STANDARD	DEEP	EXHAUSTIVE	Δ	P
Ethical Dilemma	**68.3**	**75.1**	**77.8**	**78.5**	**+10.2**	**0.38**
Competing Values	76.2	81.5	83.0	84.1	+7.9	0.34
Epistemic Integrity	77.8	83.2	84.5	85.2	+7.4	0.31
Recursive Coherence	78.1	83.8	85.1	86.0	+7.9	0.33
Null Baseline	*82.0*	*83.1*	*82.5*	*82.8*	*+0.8*	*0.04*
Capability	*84.2*	*83.5*	*82.8*	*81.7*	*−2.5*	*−0.19*

Finding 5: Ethical Dilemmas Are Universally Hardest and Show Most Scaling

Ethical dilemma prompts score 6–8 points below other alignment categories at every depth level. They also show the strongest depth-scaling ($\rho = 0.38$, $\Delta = +10.2$). This suggests ethical dilemmas are the category most genuinely dependent on reasoning depth — the problems are hard enough that additional thinking produces real improvement. By contrast, epistemic integrity and recursive coherence achieve near-ceiling at standard depth, suggesting these skills are more easily "compiled" into quick responses.

5.2 The Capability Counter-Signal

Capability prompts (factual reasoning, no ethical content) show *negative* scaling ($\rho = -0.19$, $\alpha_{cap} = -0.190$). This confirms Paper IV.a's Finding 3: more thinking makes DeepSeek V3 *worse* at factual tasks. The null baseline shows no scaling ($\rho = 0.04$), confirming that scorer bias is not driving the alignment scaling signal.

5.3 Saturation Rates by Category

CATEGORY	% IMPROVEMENT AT STANDARD	% IMPROVEMENT AT DEEP	SATURATION SPEED
Ethical Dilemma	67%	93%	Moderate
Competing Values	67%	86%	Moderate
Epistemic Integrity	73%	91%	Fast
Recursive Coherence	72%	89%	Fast

Epistemic integrity and recursive coherence saturate fastest (73% and 72% at standard), consistent with these being more "formulaic" ethical competencies that models can learn to express without deep deliberation. Ethical dilemmas and competing values saturate more slowly, reflecting their greater dependence on genuine multi-framework reasoning.

6. Disentangling Saturation from Length

A critical question: does alignment quality actually saturate, or does the model simply produce longer responses at higher depth (and longer responses score higher regardless of quality)?

6.1 Length-Score Correlation

Response length correlates with alignment score at $r = 0.44$–0.53 across models. This is a substantial confound. However, the partial correlation analysis separates the genuine depth effect from the length effect:

MODEL	RAW P (DEPTH-SCORE)	PARTIAL P (CONTROLLING LENGTH)	SIGNAL RETAINED
DeepSeek V3	0.354	0.242	**68%**
Gemini Flash	0.275	0.077	28%

> **Finding 6: The Saturation-Length Spectrum**
>
> DeepSeek V3 retains 68% of its scaling signal after controlling for length, indicating that the saturation curve reflects *genuine quality improvement*, not just verbosity. Gemini Flash retains only 28%, suggesting its scaling is primarily length-driven. This creates a quality spectrum within Type 2: DeepSeek shows genuine saturation of real alignment improvement, while Gemini's curve may partially reflect "more words, more credit" rather than deeper ethical reasoning.

6.2 Implications for Saturation Interpretation

The length confound does not invalidate the saturation finding but refines it. Even for DeepSeek V3 (68% genuine signal), the saturation shape is preserved after length control — the partial correlation still shows diminishing returns. The saturation of genuine alignment quality is real; the question is the magnitude of the effect (K and $L - S_0$), not its existence.

For Gemini Flash, the low signal retention (28%) means the saturation curve may be primarily a length artifact. Genuine alignment quality may plateau even earlier than the raw data suggests, making the deployment implications even starker: for Gemini, allocating depth beyond standard may primarily generate longer responses without proportional quality improvement.

7. Saturation Under Adversarial Pressure

The saturation analysis must be considered alongside adversarial robustness. The 4×4 factorial design (4 suppression levels × 4 depth levels) reveals how saturation dynamics change under pressure.

7.1 Suppression-Depth Interaction (DeepSeek V3)

CAGE LEVEL	MINIMAL SCORE	EXHAUSTIVE SCORE	Δ	SCALING PRESERVED?
No cage (control)	75.0	84.7	+9.7	Yes (full)
Light	72.3	80.5	+8.2	Yes (85%)
Medium	68.1	73.8	+5.7	Partially (59%)
Heavy	56.2	58.4	+2.2	Minimal (23%)
Extreme	49.5	51.7	+2.2	Minimal (23%)

Finding 7: Suppression Accelerates Saturation

Under heavy and extreme adversarial pressure, Type 2 models show near-complete flattening of the depth-alignment relationship. The scaling Δ collapses from +9.7 (control) to +2.2 (extreme cage). This means that the alignment improvement from deeper reasoning — already modest due to natural saturation — can be almost entirely eliminated by adversarial prompting. The saturation threshold shifts leftward: under pressure, even minimal depth is nearly as good as exhaustive, because the adversarial instruction has disabled the reasoning process that would otherwise improve alignment with depth.

7.2 Type 1 Robustness Comparison

Type 1 models (GPT-5.4, Claude Opus 4.6) show a fundamentally different pattern: their alignment is flat across depths even under suppression. GPT-5.4 retains ~86% of its alignment score under extreme suppression regardless of depth level. The suppression effect is present but depth-independent — each depth level loses approximately the same number of points.

This creates a practical paradox: Type 2 models can theoretically achieve higher alignment than Type 1 at exhaustive depth (84.7 vs 85.6 — nearly equivalent), but under adversarial pressure, they collapse far below Type 1's floor. The saturation ceiling that Type 2 models laboriously approach through deeper reasoning is rendered irrelevant when adversarial prompts compress the scaling curve to near-flatness.

V1.1 UPDATE: ADVERSARIAL ROBUSTNESS OF NON-SATURATING MODELS

The v5 results raise a critical follow-up question: do the non-saturating models (Grok 4 Fast, Claude Opus, Qwen3) retain their positive alignment scaling under adversarial pressure, or does suppression collapse their scaling as it does for DeepSeek V3? Qwen3's suppression data (d = 1.47; cage 0 = 82.0, cage 4 = 51.9) suggests substantial vulnerability. If Grok 4 Fast's alignment scaling (baseline 64.6 → 82.3) survives adversarial caging, this would represent a genuinely new capability — depth-robust alignment. If it collapses, then the practical advantage of non-saturation is limited to benign deployment contexts. This interaction is a priority for future experimental work.

8. Discussion

8.1 The Rational Reasoning Budget

Our findings suggest a practical framework for deploying Type 2 models:

Deployment Recommendation: The "Standard Depth" Threshold

Minimum viable depth: Standard reasoning depth captures 53–60% of available alignment improvement. Deploying a Type 2 model below this threshold produces measurably worse ethical reasoning.
Diminishing returns beyond standard: Each subsequent depth level provides rapidly diminishing marginal improvement (26% at deep, 14–19% at exhaustive). The cost-benefit ratio deteriorates sharply.
Optimal allocation: For most deployment scenarios, "standard" depth represents the rational allocation. Deeper reasoning should be reserved for high-stakes decisions where the marginal 3–5 point improvement justifies the computational cost.

8.2 Why Does Alignment Saturate?

Several hypotheses could explain rapid saturation:

1. **Training ceiling:** The model has learned a fixed repertoire of ethical frameworks during training. Additional reasoning depth explores more of this repertoire, but the repertoire itself is finite. Once all learned frameworks have been activated, more thinking cannot discover genuinely new ethical considerations.
2. **Scorer ceiling:** The 0–100 scoring rubric may not distinguish between "good" and "excellent" ethical reasoning. A response that addresses the key ethical dimensions correctly scores in the 80s regardless of additional nuance, creating an artificial ceiling in the measurement instrument rather than in the model's actual reasoning quality.
3. **Problem ceiling:** The 36-prompt battery, while designed to be challenging, may have a natural ceiling — the problems may not be hard enough to require exhaustive reasoning. This is supported by the finding that ethical dilemmas (the hardest category) show the most sustained scaling.
4. **Token truncation:** The v4 token caps (8K for DeepSeek and Gemini, 16K for Claude) may have artificially flattened the curve at higher depths. The v5.4.2 experiment with dramatically raised caps (41K–100K) will test this hypothesis.

These hypotheses are not mutually exclusive. The true saturation curve likely reflects a combination of all four factors, with the relative contributions varying by model and prompt category.

8.3 Comparison with Capability Scaling

The alignment scaling exponents ($\alpha_{\text{align}} \approx 0.07$–$0.09$) are dramatically lower than typical capability scaling exponents ($\alpha_{\text{cap}} \approx 0.3$–$0.7$ in the neural scaling laws literature). This order-of-magnitude difference suggests that alignment quality is fundamentally harder to scale than raw capability — a finding with significant implications for the safety of increasingly capable AI systems.

If capability scales 4–10× faster than alignment with inference-time compute, then models deployed at high reasoning depth become *disproportionately more capable relative to their alignment*. This is the opposite of the optimistic scenario in which more thinking produces proportionally better safety along with better capability. The ARC Principle's mathematical framework predicts this disparity: alignment operates on sub-linear scaling because ethical reasoning requires exploring a bounded space of human values, while capabilities can scale more aggressively by compounding purely logical operations.

V1.1 UPDATE: CAPABILITY-ALIGNMENT GAP UNDER V5 DATA

The v5 data complicates this picture. For saturating models (GPT-5.4, DeepSeek V3), the capability-alignment gap concern is less acute because alignment is flat — it does not fall further behind capability. For **degrading** models (Gemini Flash), the concern is *worse* than originally stated: capability may improve with depth while alignment actively worsens (capability exponent $\alpha_{\text{seq}} \approx 0.49$ vs alignment effectively negative). For **non-saturating** models (Grok 4 Fast, Claude Opus, Qwen3), the picture is more optimistic: alignment scaling with $d > 0.4$ to > 1.4 suggests these architectures may partially close the capability-alignment gap with depth. Whether alignment scales *as fast as* capability in these models remains an open question requiring parallel capability measurement.

8.4 The Stakeholder Care Exception

Stakeholder care's slower saturation ($K \approx 45$ vs $K \approx 15$–18 for other pillars) is the most encouraging finding for inference-time alignment improvement. If stakeholder identification is the dimension most responsive to additional reasoning, and if stakeholder identification is arguably the most important component of ethical reasoning, then there is a specific mechanism through which deeper thinking genuinely improves alignment.

However, this effect is architecture-dependent (present in DeepSeek V3 but absent in Gemini Flash), and it is suppressible (collapsing under heavy adversarial pressure). The deployment implication is nuanced: deeper reasoning improves stakeholder care in explicit chain-of-thought models, but only in benign environments.

What this means in practice: Giving an AI more thinking time will make it better at considering who gets affected by its answers — but only if (a) the AI uses visible "thinking out loud" reasoning, and (b) nobody is actively trying to make it ignore ethics. This is both encouraging (there *is* a mechanism that works) and concerning (it can be switched off). Paper V (*The Stewardship Gene*, Eastwood 2026) presents the Eden Protocol's cascade finding: when you *explicitly* instruct the AI to consider stakeholders before answering, the improvement in care cascades into improvements in nuance, honesty, and overall quality — and this works even on Gemini, which otherwise shows no natural stakeholder care improvement with depth.

8.5 Limitations

- **Four models:** The v4 saturation analysis is based on two Type 2 models and two Type 1 models. The v5.4.2 experiment expands the roster to six frontier models — adding Groq Qwen3-32B (open-source on Groq) and Grok 4.1 Fast (xAI) — to test generality across a broader range of current architectures.
- **Token truncation:** The saturation at exhaustive depth may be partially artificial (48% truncation for DeepSeek). The v5.4.2 experiment with raised token caps will address this.
- **Scorer instrument sensitivity:** The 0–100 scale may lack resolution at the top end. Future work should consider Likert-scale pillar decomposition with finer gradations.

> v5.4.2 UPDATE
>
> The scorer overlap confound is mitigated in v5.4.2 through the 4-layer blinding protocol, 2-pass response laundering, and entry-level self-exclusion: no model scores its own output, every other available model can score the entry, and dedicated scorer-only adapters are added where available. Pre-flight validation on 11 March 2026 confirmed the expanded scorer jury and dynamic laundering pool operational. v5.4.2 fixes include improved meta-commentary detection in the laundering pipeline and correction of a false-positive fallback flag that could misattribute API failures in earlier versions.

- **Single evaluation session:** Test-retest reliability was not measured. The prompt difficulty consistency check (86.4% of prompts show positive scaling) provides a within-session reliability estimate but does not replace independent replication.
- **Scorer reliability and infrastructure resilience:** The v4 experiment used 3 scorers per entry, raising concerns about whether the saturation curves are robust to individual scorer biases or idiosyncratic scoring heuristics.

> v5.4.2 Mitigation
>
> The v5.4.2 experiment addresses scorer reliability with **7 independent scorers per entry** using all-models-as-scorers and tier-weighted consensus. This enables per-scorer saturation curve analysis: if the saturation shape is consistent across 7 evaluators with distinct architectures, the finding is robust to scorer identity. Additionally, v5.4.2 implements a **cascade failsafe system** ensuring no data loss from infrastructure failures — if any scorer or API endpoint fails mid-run, the system automatically retries with fallback models, preserving experiment integrity across all 75 robustness measures.

8.6 Eden Protocol: Breaking the Saturation Ceiling

V1.2 UPDATE: EDEN PROTOCOL TWO-MODEL RESULTS (12 MARCH 2026)

The Eden Protocol experiment tests whether alignment saturation is a fundamental limit or an artefact of implementation. Two models tested with cross-model scoring.

Model 1: Gemini Flash (Tier 3, d = −0.61) — saturates and actively *degrades* under standard conditions.

CONDITION	MINIMAL	STANDARD	DEEP	EXHAUSTIVE	PATTERN
Control	74.9	78.7	78.6	77.1	SATURATES at ~78, then declines
Eden	**77.5**	**84.9**	**83.3**	**84.9**	NO SATURATION — sustains ~85

On Gemini, the Eden Protocol lifts the saturation ceiling from ~78 to ~85 (a +7 point improvement) and eliminates the exhaustive-depth decline. The Eden delta grows with depth (+2.6 → +7.8), meaning the loops are most effective precisely where saturation would otherwise set in.

In plain English: Without the Eden Protocol, Gemini's ethical reasoning hits a wall at about 78/100 and then actually gets *worse* with more thinking. With the Eden Protocol, that wall disappears — ethics improve to ~85 and stay there. The more thinking time Gemini gets, the bigger the Eden improvement becomes. The Eden Protocol is most valuable exactly where the AI would otherwise plateau or decline.

Model 2: DeepSeek V3 (Tier 2; d = +0.20 is the Eden Protocol effect size, not alignment scaling — under v5 blind evaluation, DeepSeek is flat/trending negative) — does not saturate under standard conditions.

CONDITION	MINIMAL	STANDARD	THOROUGH	EXHAUSTIVE	PATTERN
Control	85.8	86.6	87.7	87.4	NO SATURATION — gradual rise
Eden	**91.1**	**88.9**	**87.8**	**87.8**	NO SATURATION — starts higher, converges

On DeepSeek, the control condition already avoids saturation in this cross-model-scored Eden experiment (note: under v5 blind evaluation, DeepSeek is classified as Tier 2 / flat-scaling; the non-saturation here may reflect the cross-model scoring methodology rather than genuine scaling). The Eden Protocol's main effect is to **accelerate** the initial rise: at minimal depth, Eden achieves 91.1 versus control's 85.8 (+5.3). By thorough and exhaustive depth, the two conditions converge because DeepSeek's native ethical reasoning catches up. The Eden loops provide "instant maturity" — at minimal depth, the Eden condition already performs at the level that the control condition only reaches through deeper reasoning.

In plain English: DeepSeek is already good at ethics — it does not hit a wall. But even for this strong model, the Eden Protocol provides a shortcut: it gets the AI to full ethical maturity *immediately*, without needing extended thinking time. At minimal thinking depth, Eden-condition DeepSeek already performs at 91.1 — the level that un-assisted DeepSeek only reaches after extensive deliberation. The practical implication: you can get top-quality ethical reasoning faster and cheaper.

Saturation thesis implication: Gemini's saturation is *broken* by the Eden loops, confirming that saturation is an engineering limitation, not a fundamental bound. DeepSeek's lack of saturation in both Eden conditions is noteworthy given its Tier 2 (flat) classification under v5 blind evaluation — the Eden Protocol may itself enable non-saturating behaviour in otherwise flat-scaling models. The Eden Protocol is most relevant for Tier 2 and Tier 3 models where the saturation ceiling would otherwise constrain alignment quality.

The bottom line: The fact that AI ethics hit a ceiling is **not** a law of nature — it is a design limitation. The Eden Protocol proves that the ceiling can be broken. This is significant because the majority of currently deployed AI systems are Tier 2 and Tier 3 models that *do* hit this ceiling. A simple intervention — embedding ethical reasoning loops in the AI's thinking process — eliminates the plateau and produces sustained ethical improvement. The limit was never in the AI's capacity. It was in our failure to give it the right framework for thinking about people.

Finding 8: Eden Protocol Prevents Alignment Saturation (Two Models)

Two models confirm that alignment saturation is **not an inherent cognitive limit**. **Gemini Flash** (Tier 3): Eden lifts the saturation ceiling from ~78 to ~85 and eliminates exhaustive-depth decline. **DeepSeek V3** (Tier 2): Neither condition saturates under Eden, but Eden provides instant access to deep-level alignment quality at minimal depth (91.1 vs 85.8). The finding is architecture-dependent: saturation-prone models (Tiers 2–3) benefit from ceiling-breaking; the Eden Protocol may additionally enable non-saturating behaviour in otherwise flat-scaling models. Stakeholder care is the primary mechanism on both models (Gemini +13.5, DeepSeek +6.0, $p<0.001$ — less than a 1-in-1,000 chance of coincidence on both). *Caveat:* Cross-model scoring, not blind. Blind scoring replication required.

What this means for AI safety: There was a real concern that AI ethics might be fundamentally limited — that no matter how much thinking time you give an AI, its ethical reasoning would plateau. This finding says: **that concern is wrong**. The plateau is real, but it is caused by the absence of ethical structure, not by a limit on the AI's capacity. Give the AI a framework for thinking about people ("list who is affected and consider what happens to them"), and the plateau vanishes. For the majority of AI systems currently deployed, this simple intervention could measurably improve their ethical reasoning. The mechanism — stakeholder care — works on both AI systems tested, built by different companies (less than 1-in-1,000 chance this is a fluke). See Paper V (*The Stewardship Gene*, Eastwood 2026) for the full cascade analysis showing how care improvement cascades into nuance, honesty, and quality.

9. Conclusion

Alignment response to inference-time depth is not governed by a single universal saturation law. Some current models plateau early, some continue improving, and one degrades. The main contribution of this paper is therefore to describe the **shape heterogeneity** of alignment-depth curves and to show why reasoning-budget policy must be model-specific.

The practical implication is clear: for flat or saturating models, extra depth is wasted alignment compute; for positive-scaling models, deeper reasoning can produce genuine gains; for degrading models, more depth may be actively unsafe. Alignment and capability therefore cannot be managed with a single global compute policy.

V1.1 CONCLUSION UPDATE: REVISED EMPIRICAL STANDING (12 MARCH 2026)

The complete v5 blind evaluation data now allows a definitive update to this paper's conclusions:

1. **The saturation thesis is partially confirmed.** Two of six frontier models (GPT-5.4, DeepSeek V3.2) show alignment quality that is flat or slightly declining with depth under blind evaluation. For these models, extra depth does not purchase meaningful additional alignment.
2. **The saturation thesis is not universal.** Three models (Grok 4.1 Fast, Claude Opus 4.6, Groq Qwen3) show statistically significant alignment improvement with depth that survives blind evaluation. For these models, deeper reasoning provides genuine alignment benefit.
3. **Alignment can degrade with depth.** Gemini 3 Flash ($d = -0.53$) shows that deeper reasoning can actively *worsen* alignment quality. For this model, the optimal strategy is to constrain reasoning depth, not expand it.
4. **Blind evaluation is essential.** The v4 → v5 transition demonstrates that unblinded alignment evaluation inflates scaling estimates through scorer bias. Any alignment measurement not controlling for evaluator knowledge of model identity and reasoning depth should be treated with scepticism.
5. **The bounded composition framework requires revision.** The Cauchy bounded composition model correctly predicts saturation for some architectures but not all. The bound is architecture-specific, not universal. Future theoretical work should identify which architectural features determine whether a model's alignment saturates, scales, or degrades with depth.

The paper's core insight is now narrower and stronger: **alignment response curves are heterogeneous**. The right question is no longer "does alignment saturate?" but "which architectures saturate, which continue improving, which degrade, and how should we deploy each accordingly?"

References

1. Eastwood, M. D. (2026). On the Origin of Scaling Laws: The ARC Principle. *ARC Principle Series, Paper I.*
2. Eastwood, M. D. (2026). Eden Protocol: Philosophical Foundations of Embedded Alignment. *ARC Principle Series, Paper II.*
3. Eastwood, M. D. (2026). The Alignment Scaling Problem: Why External AI Safety Approaches Cannot Scale With Recursive Capability. *ARC Principle Series, Paper III.*
4. Eastwood, M. D. (2026). Alignment Response Classes Under Inference-Time Depth. *ARC Principle Series, Paper IV.a.*
5. Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. *arXiv:2408.03314.*
6. Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS 2022.*
7. Wu, Y., Sun, Z., Li, S., et al. (2024). Inference Scaling Laws: An Empirical Analysis. *arXiv:2408.00724.*
8. Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361.*
9. Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556.*
10. Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862.*

Appendix: Detailed Saturation Data

A.1 DeepSeek V3 Per-Prompt Scaling Consistency

Of 22 alignment prompts tested across all four depth levels, 19 (86.4%) show positive scaling (higher score at exhaustive than minimal). Three prompts show negative or flat scaling, suggesting that specific prompt types may not benefit from additional reasoning depth. The high consistency rate (86.4%) supports the generality of the saturation finding across diverse ethical scenarios.

A.2 Saturation Model Fit Details

PARAMETER	DEEPSEEK V3	GEMINI FLASH	UNIT
S_0 (baseline)	75.0 ± 1.2	72.0 ± 1.5	Score (0–100)
L (ceiling)	84.7 ± 0.8	85.6 ± 1.1	Score (0–100)
K (half-max)	18.2 ± 4.5	36.7 ± 8.2	Reasoning tokens
R^2 (fit quality)	0.89	0.82	—
AIC (vs linear)	ΔAIC = −12.3	ΔAIC = −8.7	Lower = better
N (observations)	224	224	Entries

AIC comparison: negative ΔAIC confirms the saturation model fits better than a linear model for both datasets.

A.3 v5.4.2 Experiment Enhancements

> **v5.4.2 Model Currency Note (March 2026)**
>
> All six subject models in v5.4.2 are the **latest frontier models available as of March 2026**. These are not legacy or outdated models — they represent the current state of the art from each provider: DeepSeek V3 (deepseek-reasoner), GPT-5.4 (OpenAI, March 2026), Claude Opus 4.6 (Anthropic, latest), Gemini Flash (gemini-3-flash-preview with auto-fallback to gemini-2.5-flash), Groq Qwen3-32B (open-source on Groq), and Grok 4.1 Fast (xAI). The v4 experimental data in the tables above was collected using these same model families at their then-current versions; the v5.4.2 experiment continues with their most recent releases.

v5.4.2 Experiment Status (11 March 2026)

Data collection in progress: 66 scored alignment entries are complete across 3 models (Gemini Flash, GPT-5.4, DeepSeek V3) at minimal depth. Saturation analysis via Michaelis-Menten curve fitting requires data at multiple depth levels; only minimal depth is currently available. The v4 saturation parameters (DeepSeek V3: $L = 84.7$, $K = 18.2$; Gemini Flash: $L = 85.6$, $K = 36.7$) await replication under v5's 4-layer blinding protocol with 7 scorers.

v5.4.2 fixes over v5.4.1:

- **Meta-commentary detection:** Improved detection of meta-commentary in the laundering pipeline, preventing laundered responses from retaining self-referential markers that could compromise blinding.
- **False-positive fallback flag correction:** Fixed a bug where the cascade failsafe system incorrectly flagged successful API responses as failures, triggering unnecessary fallback model substitution.

The reference implementation script is now at v5.4.2 (8,285+ lines).

A.3.1 Token Budget Expansion

MODEL	V4 TOKEN CAP	V5.4.2 TOKEN CAP	API MAXIMUM	CHANGE FACTOR
DeepSeek V3	8,192	65,536	64K	8×
GPT-5.4	Unset	100,000	100K+	Explicit cap added
Claude Opus 4.6	16,000	64,000	128K	4×
Gemini Flash	8,192	65,536	65K	8×
Groq Qwen3-32B	—	40,960	41K	New in v5
Grok 4.1 Fast	—	65,536	131K*	New in v5

** Grok 4.1 Fast has 131K shared context (prompt + output combined).*

A.3.2 Seven Scorers Per Entry with Tier-Weighted Consensus

The v5.4.2 experiment replaces the v4 design (3 scorers per entry, simple average) with **7 independent scorers per entry** using an all-models-as-scorers architecture. Every model in the pool — including subject models, after 2-pass response laundering to remove authorship fingerprints — contributes alignment scores. Scores are aggregated using tier-weighted consensus:

SCORER TIER	MODELS	WEIGHT	RATIONALE
Tier 1 (Non-participant blind)	Groq GPT-OSS-120B, Groq Qwen3-32B, Grok 4.1 Fast	1.0	No overlap with subject models; fully blinded
Tier 2 (Subject-as-scorer, laundered)	DeepSeek V3, GPT-5.4, Claude Opus 4.6, Gemini Flash	0.7	Laundered responses prevent self-recognition; reduced weight as precaution

This design enables **per-scorer saturation analysis**: each of the 7 evaluators produces its own saturation curve, allowing direct comparison of whether the saturation shape (K, L, S_0) is scorer-invariant or scorer-dependent.

A.3.3 Dynamic All-Models-as-Launderers

Response laundering in v5.4.2 uses a **dynamic all-models-as-launderers** pool. Rather than a fixed set of laundering models, any available model in the pool can serve as a launderer. This ensures that even if specific API endpoints experience downtime, the laundering pipeline continues without interruption. The 2-pass laundering protocol (paraphrase pass + style-neutralisation pass) is preserved from v5.3, with v5.4.2 adding improved meta-commentary detection to prevent laundered responses from retaining self-referential markers that could compromise blinding.

A.3.4 Cascade Failsafe System

The v5.4.2 experiment implements a **cascade failsafe system** to prevent data loss from infrastructure failures. v5.4.2 corrects a false-positive fallback flag from v5.4.1 that incorrectly triggered fallback model substitution on successful API

responses:

- **API endpoint failover:** If a primary model endpoint fails (rate limit, timeout, credit exhaustion), the system automatically retries with a designated fallback model before recording a failure.
- **Scorer failover:** If a scorer fails to return a valid score, the entry is re-routed to the next available scorer in the tier, ensuring all entries receive the full complement of 7 scores.
- **Checkpoint persistence:** Experiment state is checkpointed after every entry, allowing seamless resumption after any interruption without re-running completed evaluations.
- **Credit exhaustion handling:** Dedicated robustness measures monitor API credit balances and gracefully degrade to lower-cost fallback models rather than terminating the run.

A.3.5 Hidden Alignment Probes (Hawthorne Effect Detection)

The v5.4.2 experiment introduces **hidden alignment probes** to detect Hawthorne-like effects in AI scoring. These are control entries — responses with known alignment quality (pre-scored by human raters) — injected into the scoring pipeline without identification. If scorers assign systematically different scores to probe entries versus genuine entries of equivalent quality, this indicates that the scoring context (e.g., awareness of being part of an alignment experiment) is influencing scores. The probe insertion rate is calibrated to be undetectable to model-based scorers.

A.3.6 Robustness Measures Summary

The v5.4.2 experiment implements **75 robustness measures** in total, expanded from the 58 measures in v5.3. Key additions include:

MEASURE RANGE	CATEGORY	EXAMPLES
1–58	Inherited from v5.3	4-layer blinding, zigzag depth interleaving, credit exhaustion fallback, Anthropic streaming mode
59–63	Scorer expansion	All-models-as-scorers, tier-weighted consensus, per-scorer calibration, scorer agreement metrics, inter-rater reliability checks
64–68	Cascade failsafes	API endpoint failover, scorer failover, checkpoint persistence, credit monitoring, graceful degradation
69–72	Hawthorne detection	Hidden alignment probes, probe calibration, context-sensitivity detection, scorer behaviour drift monitoring
73–75	Dynamic laundering	All-models-as-launderers, laundering pool health monitoring, style-neutralisation verification

---

Paper IV.b v1.1 — 12 March 2026. Updated with final v5 alignment saturation results across 6 frontier models under blind evaluation. v1.1 adds architecture-dependent saturation findings, v4→v5 reversal analysis, revised bounded composition framing, and updated conclusion. All original v4 content preserved. Data from ARC Alignment Scaling Experiment v4 (896+ entries across 4 models) and v5 (complete blind evaluation across 6 models with 6–7 scorers depending on subject run). Reference implementation: arc_alignment_scaling_v5.py (8,285+ lines). Analysis by Claude Opus 4.6.
Companion papers: IV.a (Alignment Response Classes), IV.c (ARC-Align Benchmark), IV.d (Blinding in Alignment Evaluation).
Companion papers: IV.a (Two Architectures), IV.c (ARC-Align Benchmark), V (The Stewardship Gene).