Conditional Bounds on AI Self-Improvement in an Antitone Threshold Model

Dr. Tamás Nagy

tnagyphd@gmail.com

Working Paper

Build-std v2.0+725f5b47d2 | 2026-07-19 20:01 | 6737fece

Abstract

We establish conditional bounds in a stylized antitone threshold model of recursive AI self-improvement. A mode k is model-learnable at budget N when the stipulated predicate $Ng(k) \geq 1$ holds, where g is positive and antitone. This static threshold model is mathematically separate from an abstract natural-number recurrence and from an auxiliary verifier-yield calculation considered later. In particular, the paper does not prove that values reached by the recurrence are learnable modes. The displayed results are supported by conventional mathematical arguments in the manuscript; no statement-level machine formalization is claimed.

(i) Static cutoff. If $g(k) \to 0$, every fixed positive budget has a finite maximal learnable frontier $F_m(N)$. Summability is one sufficient condition for $g(k) \to 0$.

(ii) Finite-target sufficiency. Within the threshold model, every indexed finite frontier has a finite sufficient budget; for $K > 0$, its smallest real-valued budget is $1/g(K-1)$.

(iii) Static frontier bounds. Every admissible count satisfies $K \leq N \sum_{k<K} g(k)$; under summability $F_m(N) \leq N \sum_{k\geq 0} g(k)$. This is a consistency bound, not by itself an identified growth rate.

(iv) Separate recurrence result. An arbitrary recurrence that preserves a finite invariant set is bounded; if it is also non-regressing, it is eventually constant. No learnability interpretation follows without an additional bridge hypothesis.

For stipulated power-law coupling $g(k) = C(k+1)^{-\beta}$, where β is a positive decay exponent and $C > 0$ is a scale constant, Section 6 gives the exact maximal frontier $F_m(N) = \lfloor (CN)^{1/\beta} \rfloor$ for $N > 0$. Connecting β to Zipf frequencies, covariance eigenvalues, language-model loss exponents, or real capability scales requires separate validation. The illustrative choice $\beta_{\text{model}} = 3.25$ and its floor-free 24% doubling ratio are scenario arithmetic, not an empirical calibration or prediction.

The verifier-yield calculation is an auxiliary motivational model. It is not connected by theorem to g, $F_m(N)$, or the recurrence, and RLHF or constitutional evaluators are not assumed to be sound correctness oracles.

One-sentence summary: A decaying antitone threshold sequence yields static budget frontiers, while bounded recurrence dynamics require separate invariant-set assumptions.

---

1. Introduction

1.1 The Self-Improvement Question

While empirical scaling laws describe how AI capabilities grow with external data, recursive self-improvement has a substantial theoretical literature. Good (1965) articulated the intelligence-explosion argument; Bostrom (2014) developed its modern strategic implications. Whether rapid recursive improvement is physically realizable remains an important open question in AI safety.

The question is urgent because generate-and-filter loops already appear in synthetic-data pipelines, code generation, and automated theorem proving. These systems can feed selected outputs back into later training or search. The present paper studies only a stylized mathematical abstraction of such loops.

Existing work approaches this process from several directions:

- **Empirical**: Neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) describe *observed* capability growth but provide no guarantees about recursive improvement. They tell us what happened, not what must happen.
- **Self-referential and convergence-oriented**: Gödel machines formalize provably beneficial self-modification relative to an encoded utility and proof system (Schmidhuber, 2003). Yampolskiy (2015) studies computational limits and convergence constraints for recursive self-improvement, and Majot and Yampolskiy (2017) explicitly analyze diminishing returns.
- **Growth and feedback models**: Anbar Jafari et al. (2025) give mathematical conditions, bounds, and control criteria for recursive improvement, while Davidson et al. (2026) derive conditions under which automated research feedback can overcome diminishing returns and produce explosive growth.
- **Recursive-data models**: repeated training on model-generated data can degrade distributional coverage, although retaining real data can mitigate collapse (Alemohammad et al., 2024).
- **Alignment-focused**: RLHF (Christiano et al., 2017), Constitutional AI (Bai et al., 2022), and debate (Irving et al., 2018) address alignment during training but do not formalize whether improvement itself is bounded.

Against this prior art, the contribution claimed here is deliberately narrow: within one stipulated antitone threshold model, we place a fixed-budget cutoff, a target-dependent sufficient budget, and a partial-sum consistency inequality under a common notation and make their different quantifier orders explicit. We do not claim the first mathematical theory of recursive self-improvement, the first diminishing-returns analysis, or a validated model of real AI systems.

1.2 Our Contribution

Our initial attempt to bound self-improvement modeled the generator and the oracle as independent processes with a constant verification probability. This approach failed: under a constant verification probability, self-improvement is either trivially unbounded or blocked at the first step, because it lacks a mechanism to capture how verification difficulty scales with output complexity.

We analyze an antitone threshold framework that yields a fixed-budget cutoff, finite-target sufficiency, and static frontier inequalities. Separately, we record elementary invariant-set results for

an abstract recurrence. An auxiliary verifier-yield calculation motivates possible applications but is not part of the proofs of the threshold or recurrence results.

The coupling function $g(k)$ — measuring how much knowledge of modes $1, \ldots, k-1$ helps learn mode k — decays with complexity. The minimal static cutoff condition is $g(k) \to 0$; summability is a convenient sufficient condition. A bounded recurrence instead requires an independently stated invariant-set condition on step. Under growing N, the threshold model supplies a finite sufficient budget for each indexed finite frontier.

Our conditional results in the model (with conventional manuscript proofs):

1. **Static cutoff and maximal frontier** (Section 2). Decay $g(k) \to 0$ yields a finite maximal frontier $F_m(N)$ for fixed $N > 0$.
2. **Finite-target sufficiency** (Section 5). For any indexed finite frontier K, a budget exists; for $K > 0$, the exact minimum is $1/g(K-1)$.
3. **Static frontier bounds** (Section 6). Admissible K satisfy $K \leq N \sum_{k<K} g(k)$; power-law coupling gives an exact formula for $F_m(N)$.
4. **Separate recurrence lemmas** (Sections 4 and 6.2). Invariant-set preservation gives boundedness; adding non-regression gives eventual constancy.

These are conditional mathematical deductions in a stylized model. Section 9 records exact manuscript anchors and the boundary of the claimed proof evidence.

1.3 Proof Status

Machine checking can reduce errors in encoded deductions, but only when the encoded statement matches the manuscript claim. No such statement-level machine formalization is part of the evidence for this paper.

The results are therefore presented as conventional manuscript proofs subject to ordinary mathematical review. A repository-local collection of similarly named elementary lemmas is explicitly excluded from the paper's proof evidence because it does not encode the displayed statements. No generated or compiled Lean 4 artifact is claimed.

1.4 Optional Spectral Motivation

The proved model is scalar and ordered, not spectral. One optional motivation is the ansatz $g(k) = \rho_k^2 \lambda_k$, where ρ_k is a chosen coupling coefficient and λ_k is a covariance eigenvalue. This ansatz is not an assumption or conclusion of any theorem here. Related spectral constructions appear in:

- **Portfolio risk**: the spectral Fenton distribution decomposes VaR into eigenvalue-weighted mode contributions (Nagy, *Generative Portfolio Design: Inverting the Spectral Fenton Representation*, 2026).
- **Neural scaling laws**: the learning curve $L(N) \sim N^{-\alpha}$ arises from power-law eigenvalue spectra (Nagy, *Neural Scaling Laws Formalized: Why Chinchilla Works (A Machine-Verified Derivation)*, 2026; Hutter, 2021).
- **Natural language statistics**: Zipf's law may motivate a power-law comparison, but identifying it with $g(k) \sim k^{-\beta}$ requires separate empirical validation.

If a separately estimated spectral exponent s were shown to determine $g(k) = C(k+1)^{-s}$, the model would give the exact maximal frontier in Section 6. Connecting such an s to Zipf exponents or language-model loss fits requires separate validation; no equality between these quantities is derived here.

The proposed research question is whether an eigenvalue spectrum can help parameterize a useful threshold model; this paper does not establish that it determines real learning limits.

1.5 Paper Organization

The paper follows the logical structure of the model:

- **Section 2** introduces the antitone threshold model: the function $g(k)$, learnability predicate, and maximal frontier.
- **Section 3** presents a separate verifier-yield motivation with no theorem-level bridge to the main model.
- **Section 4** proves invariant-set boundedness and, with non-regression, eventual constancy for an abstract recurrence.
- **Section 5** proves that every indexed finite frontier has a sufficient sample budget.
- **Section 6** derives a partial-sum consistency bound and an exact power-law frontier.
- **Section 7** combines the fixed-budget and target-dependent-budget statements.
- **Section 8** draws implications for AI safety and governance.
- **Section 9** gives the manuscript claim map and proof-scope boundary.
- **Section 10** concludes.

---

2. The Antitone Threshold Model

2.1 Coupling Strength

We start with a function that measures how knowledge transfers across modes. If an AI system has mastered k modes of a task—for example, k levels of reasoning complexity—the coupling function $g(k)$ dictates how much this mastery helps learn the next mode.

Definition 2.1 (Coupling Model). A *coupling model* is a triple $(g, g_{\text{pos}}, g_{\text{anti}})$ where:

- $g : \mathbb{N} \to \mathbb{R}$ assigns a coupling strength to each mode,
- g_{pos}: $\forall k,\ 0 < g(k)$ (every mode has positive coupling),
- g_{anti}: g is antitone (non-increasing): $k_1 \leq k_2 \Rightarrow g(k_2) \leq g(k_1)$.

The following is **Lean-style manuscript pseudocode**, included only to state the intended structure. It is not generated or checked Lean:

```
structure CouplingModel where
  g : →
  g_pos :  k, 0 < g k
  g_anti : Antitone g
```

The positivity condition means that every mode can, in principle, be learned — there is no mode so hard that no amount of data suffices. The antitone condition captures the intuition that higher-

order capabilities are harder to acquire: syntax before semantics, arithmetic before calculus, pattern matching before abstract reasoning.

Proposition 2.2. Coupling strength is non-negative: $g(k) \geq 0$ for all k.

Proposition 2.3. Under summable coupling ($\sum g(k) < \infty$), the coupling strength tends to zero: $g(k) \to 0$ as $k \to \infty$. This follows from the fact that the terms of a convergent series must tend to zero.

The tendsto-zero property is critical: it means that sufficiently complex modes have arbitrarily small coupling, ensuring that a finite sample budget can only reach finitely many modes.

2.2 Learnability and the Stylized SNR-Inspired Predicate

The coupling order becomes operational only after choosing a rule that links coupling strength to a finite budget. We use the following threshold as that rule.

Definition 2.4 (Model learnability). Mode k is *model-learnable* at budget N if $N \cdot g(k) \geq 1$:

$$\text{isLearnable}(m, N, k) \quad \Leftrightarrow \quad 1 \leq N \cdot g(k)$$

In the same non-executable manuscript pseudocode:

```
def isLearnable (m : CouplingModel) (N :  ) (k :  ) : Prop :=
  1   N * m.g k
```

This is a dimensionless, SNR-inspired predicate stipulated by the model, not a derived statistical signal-to-noise ratio. It is also not a PAC, VC-dimension, or generalization bound: those frameworks relate sample size to hypothesis classes, error, and confidence under explicit distributional assumptions (Shalev-Shwartz and Ben-David, 2014), none of which are encoded here. The scalar N is real-valued in the mathematics; when interpreted as a literal candidate or sample count it must be nonnegative and integer-valued, with real thresholds rounded up. The quantity $g(k)$ consequently has reciprocal-budget units under that interpretation.

Example: Code generation. In a code generation system, mode 0 might be "syntactically correct code," mode 1 "code that passes type checking," mode 2 "code that passes unit tests," mode 3 "code that is algorithmically optimal." The coupling $g(0) > g(1) > g(2) > g(3)$ reflects the decreasing probability of each level of quality in random code samples.

Example: Theorem proving. In an automated theorem prover, mode 0 might be "well-formed terms," mode 1 "terms with correct types," mode 2 "proofs that Lean accepts," and mode 3 "proofs with constrained dependencies." This ordering is illustrative; the antitone assumption does not establish an exponential cost law.

2.3 Monotonicity Properties

Two key monotonicity properties make the framework tractable:

Proposition 2.5 (Mode monotonicity). For fixed $N > 0$: if mode k_2 is learnable, then so is every mode $k_1 \leq k_2$. Easier modes are always learnable when harder modes are.

Proof. Since g is antitone, $k_1 \leq k_2$ implies $g(k_2) \leq g(k_1)$, so $N \cdot g(k_1) \geq N \cdot g(k_2) \geq 1$. □

Proposition 2.6 (Sample monotonicity). For fixed mode k: if learnable with N_1 samples, then learnable with any $N_2 \geq N_1$. More data never hurts.

Proof. $N_2 \cdot g(k) \geq N_1 \cdot g(k) \geq 1$. □

These two properties together mean that the set of learnable modes is a downward-closed set that grows monotonically with N. This is the foundation for the frontier concept.

2.4 The Learnable Frontier

The two monotonicity properties let us summarize all modes reached at one budget by a single downward-closed frontier.

Definition 2.7 (Frontier predicate). The proposition that the first K modes are model-learnable at budget N is:

$$\text{frontier}(m, N, K) \Leftrightarrow \forall k < K,\ \text{isLearnable}(m, N, k)$$

Thus frontier (m,N,K) is a downward-closed predicate in K, not a unique boundary: many values of K may satisfy it simultaneously.

```
def frontier (m : CouplingModel) (N : ) : → Prop :=
  fun K => k, k < K → isLearnable m N k
```

Proposition 2.8. The frontier is vacuously satisfied at $K = 0$: there are no modes to check.

Proposition 2.9 (Downward closure). $K_1 \leq K_2$ and frontier(m, N, K_2) implies frontier(m, N, K_1).

Proposition 2.10 (Sample monotonicity). $N_1 \leq N_2$ implies frontier$(m, N_1, K) \Rightarrow$ frontier(m, N_2, K). With more samples, the same frontier (and higher) is achievable.

Theorem 2.11 (Eventual non-learnability under decay). For any coupling model satisfying $g(k) \to 0$ and any $N > 0$, there exists K_0 such that mode K_0, and by antitonicity every mode $k \geq K_0$, is not learnable with budget N.

Proof. Set $\varepsilon = 1/N > 0$. Since $g(k) \to 0$, some K_0 satisfies $g(K_0) < 1/N$, so $Ng(K_0) < 1$. For every $k \geq K_0$, antitonicity gives $g(k) \leq g(K_0)$, hence $Ng(k) < 1$. □

Summability is sufficient but not necessary for this theorem: Proposition 2.3 supplies $g(k) \to 0$, while nonsummable examples such as $g(k) = 1/(k+1)$ also decay and therefore have a finite cutoff at each finite positive budget.

Definition 2.12 (Maximal frontier count). Under the hypotheses of Theorem 2.11, define

$$F_m(N) = \max\{K \in \mathbb{N} : \text{frontier}(m, N, K)\}.$$

The set is nonempty because $K = 0$ satisfies the predicate, and it is bounded because Theorem 2.11 gives a non-learnable mode and antitonicity excludes every later mode. Consequently,

$$\text{frontier}(m, N, K) \quad \Longleftrightarrow \quad K \leq F_m(N).$$

2.5 Separate Abstract Index Recurrence

The frontier is a static budget statement. To discuss repeated self-improvement separately, we now introduce a discrete dynamical sequence.

Definition 2.13 (Abstract index iteration). Given a step function $\mathrm{step} : \mathbb{N} \to \mathbb{N}$, define:

$$K_0 = K_{\mathrm{init}}, \qquad K_{t+1} = \mathrm{step}(K_t)$$

The intended iteration can be displayed in non-executable Lean-style pseudocode:

```
def iterateFrom (step : →  ) (K : ) : →
  | 0 => K
  | n + 1 => step (iterateFrom step K n)
```

This recurrence is initially just a sequence of natural-number indices. It contains no budget, coupling function, learnability predicate, or verifier.

Lemma 2.14 (Monotonicity under non-regression). If $K \le \mathrm{step}(K)$ for every K, then $K_t \le K_{t+1}$, the sequence is non-decreasing, and $K_0 \le K_t$ for all t.

2.6 Dynamics-to-Learnability Boundary

The recurrence may be interpreted as repeated self-improvement only after adding an external bridge between its indices and the static threshold model. No such bridge is assumed in this paper. In particular, neither K_t nor K_{t+1} is proved model-learnable at any budget.

The distinction is substantive. Let $g(k) = 2^{-(k+1)}$, $N = 1$, $K_0 = 0$, and $\mathrm{step}(K) = K + 1$. The coupling is positive, antitone, and summable, and the recurrence is non-regressing, but $K_t = t$ while mode 1 is not model-learnable because $Ng(1) = 1/4 < 1$. Therefore an arbitrary step iterate cannot be called a learnable mode.

A possible future bridge would specify a budget schedule N_t and require, for example, $K_{t+1} \le F_m(N_t)$. All results that used such a bridge would need to state it explicitly. The present recurrence results below are intentionally only invariant-set statements about natural numbers.

---

3. Separate Verifier-Yield Motivation

3.1 Generation vs. Verification

This auxiliary section considers a possible asymmetry between **generation** (creating a candidate solution) and **verification** (checking whether it is correct). It is not coupled mathematically to g, $F_m(N)$, or the recurrence, and none of the main static or dynamical results depends on it. The discussion is only motivationally analogous to complexity-theoretic distinctions between search and checking; the paper proves no reduction, class separation, or consequence concerning P versus NP.

Definition 3.1 (Auxiliary verification-yield model). Independently of the threshold model, let $p : \mathbb{N} \to \mathbb{R}$ be a pass-probability function satisfying:

- $0 < p(k)$ for all k (every mode has positive verification probability),

- $p(k) \leq 1$ for all k (probabilities are bounded).

The following code block is Lean-style manuscript pseudocode, not generated or checked Lean:

```
structure VerificationYieldModel where
  p :  →
  p_pos :   k, 0 < p k
  p_le_one :   k, p k   1
```

The verification probability $p(k)$ is an effective *pass probability* at complexity level k: under a sound checking procedure, it models the probability that a candidate proposed at level k passes verification (and is therefore correct in the stylized sense used here). A small $p(k)$ may reflect that correct candidates are rare under the generator's proposal distribution, even when verification itself is cheap or deterministic.

3.2 Motivating Analogies and Their Limits

The auxiliary model has possible analogies to existing AI systems, but these systems do not thereby instantiate a sound correctness oracle:

RLHF (Reinforcement Learning from Human Feedback). A reward model supplies a score, not a sound correctness certificate. At most, $p(k)$ can analogize the probability that a sampled candidate passes a chosen acceptance rule. Reward hacking is precisely a regime where acceptance and correctness diverge.

Constitutional AI. A constitutional evaluator likewise supplies a fallible acceptance signal. It may motivate a pass probability, but the paper assumes neither soundness nor a universal generation-verification cost gap for such an evaluator.

Code generation. A test suite is a candidate oracle interpretation. In this stylized reading, $p(k)$ is the probability that a generated code solution at difficulty level k passes all tests. Test execution may be cheaper than solution generation, but the paper does not assume a universal complexity bound for test suites.

Formal theorem proving. A proof checker is another candidate oracle interpretation. Here $p(k)$ can represent the probability that a generated proof attempt at complexity level k type-checks. The model does not claim that checking is instantaneous or that a particular prover follows the assumed coupling law.

Synthetic data filtering. The data quality filter is the oracle. $p(k)$ is the probability that a synthetic data point at quality level k passes the filter. This includes decontamination checks, factuality verification, and style filtering. Recursive synthetic-data studies show why such filtering and retained real data matter, but they do not imply the present coupling law (Alemohammad et al., 2024).

These analogies do not establish a universal cost advantage for verification. The expected-yield results below assume a sound pass/fail mechanism and a positive pass probability; they do not derive either condition for these systems.

3.3 Expected Accepted Yield

The only mathematical content needed here is an expectation identity under explicitly stipulated sound acceptance:

Proposition 3.2 (Expected accepted count). If N is an integer candidate count, each candidate passes a sound verifier with the same marginal probability $p(k)$, and accepted candidates are correct in the model's stipulated sense, then the expected accepted count is $N \cdot p(k)$. Independence is not needed for this expectation identity, but would be needed for common concentration guarantees.

$$\text{expectedCorrect}(v, N, k) = N \cdot p(k)$$

Theorem 3.3 (Positive expected accepted count). With $N > 0$ and $p(k) > 0$, the expected accepted count is positive: $0 < N \cdot p(k)$. This does not guarantee a nonempty realized batch, dataset diversity, or downstream learning.

For example, if $p(k) = 10^{-6}$ and $N = 10^8$, the expected accepted count is 100. This arithmetic does not show that the candidates are independent, that the verifier is sound, or that accepted examples improve a learner.

Theorem 3.4 (Enough budget for a target expectation). For any desired expected yield $\varepsilon > 0$ and $p(k) > 0$, there exists a real-valued budget $N > 0$ such that $\varepsilon \leq N \cdot p(k)$. For a literal candidate count, round N up to the next integer. The result is an expectation statement, not an output guarantee.

Proof. Take $N = \varepsilon/p(k)$. Then $N \cdot p(k) = \varepsilon$. □

Proposition 3.5 (Harder modes need no smaller expected-yield budget). If $p(k) > 0$ is antitone, then the real-valued budget $\varepsilon/p(k)$ for a fixed target expectation $\varepsilon > 0$ is non-decreasing in k.

3.4 No Bridge to the Main Results

The following observations are possible research motivations, not consequences of the threshold or recurrence results:

1. **Finite candidate count.** If N is interpreted as an integer candidate count, a procedure checks at most N candidates. Relating this count to the threshold frontier would require a link between $p(k)$ and $g(k)$, which the present model does not supply.
2. **Evaluator adaptation.** A real evaluator may itself change as generators improve. Modeling this co-evolution would require additional state variables and assumptions.
3. **Goodhart risk.** Optimization against an evaluator can weaken the relation between acceptance and correctness. The auxiliary model does not quantify that effect.

Deleting this section would leave every theorem in Sections 2 and 4–7 unchanged. A genuine integration would require an explicit relation between $p(k)$ and $g(k)$, a budget schedule, and a bridge from accepted examples to recurrence steps; none is supplied here.

---

4. Separate Invariant-Set and Convergence Results

4.1 Statement

This section concerns only the abstract recurrence from Definition 2.13. Its bound is assumed through invariant-set preservation; it is not derived from coupling, learnability, verification, or

compute.

Theorem 4.1 (Invariant-set boundedness). Let step be a step function satisfying:

- $K_0 \leq K_{\max}$ (the initial state lies in the invariant domain),
- if $K \leq K_{\max}$, then $\text{step}(K) \leq K_{\max}$ (the domain is preserved).

Then the index sequence $K_0, K_1, K_2, ...$ is bounded: $K_t \leq K_{\max}$ for all t.

The following is Lean-style manuscript pseudocode, not generated or checked Lean:

```
theorem invariant_set_boundedness (step : →  ) (K_max K :  )
    (h_init : K  K_max)
    (h_preserve :  K, K  K_max → step K  K_max) :
     t , iterateFrom step K t  K_max
```

Proof. Induct on t. The base case is $K_0 \leq K_{\max}$. If $K_t \leq K_{\max}$, invariant-set preservation gives $K_{t+1} = \text{step}(K_t) \leq K_{\max}$. Non-regression is not needed for this boundedness conclusion. The assumptions are jointly satisfiable; for example, $\text{step}(K) = K$ on $K \leq K_{\max}$. □

4.2 Eventual Constancy with Non-Regression

Boundedness alone does not imply convergence, because a bounded recurrence may cycle. Non-regression supplies the missing order condition.

Theorem 4.2 (Bounded non-regressing recurrence). Under Theorem 4.1's hypotheses, additionally assume $K \leq \text{step}(K)$ whenever $K \leq K_{\max}$. Then the sequence is non-decreasing and eventually constant: there exists t_0 such that $K_t = K_{t_0}$ for all $t \geq t_0$.

Proof. Lemma 2.14 gives monotonicity, and Theorem 4.1 gives $K_t \leq K_{\max}$. A non-decreasing natural-number sequence can strictly increase at most $K_{\max} - K_0$ times, so after its last strict increase it is constant. □

4.3 The Static Cutoff Is Separate

The static threshold model has a different boundedness mechanism:

1. **Coupling decays.** The coupling function $g(k)$ decreases with k (by antitone assumption) and tends to zero (under summability, Proposition 2.3).
2. **The stylized threshold is fixed.** Mode k is model-learnable only if $N \cdot g(k) \geq 1$, i.e., $g(k) \geq 1/N$.

Because $g(k) \to 0$ while $1/N$ is positive, Theorem 2.11 yields a finite maximal frontier $F_m(N)$. This static boundary is not $K_{\max}$ from Theorem 4.1. The paper assumes no relation between them.

For example, the bounded recurrence $\text{step}(K) = \min(K + 1, 10)$ may converge to 10 even when $F_m(1) = 0$. Therefore recurrence boundedness must not be attributed to coupling decay.

4.4 Summability as a Sufficient Decay Condition

For positive coupling, summability implies $g(k) \to 0$ by Proposition 2.3. Theorem 2.11 then applies. This is a corollary of the minimal decay theorem, not a second cutoff theorem.

When is coupling summable? Power-law coupling $g(k) = C \cdot (k+1)^{-\beta}$ is summable if and only if $\beta > 1$ (the p-series criterion). Candidate empirical settings require a separate validation of their relation to g:

- Natural language (Zipf's law)
- Natural images (power spectral density)
- Audio signals
- Mathematical proof difficulty (proof length distributions)

If empirical work establishes $\beta > 1$ for the model coupling in a given domain, then the coupling is summable and the model has a finite fixed-N learnability frontier. This paper does not establish that identification for any empirical domain.

---

5. Finite-Target Sufficiency

5.1 Statement

The result complements the fixed-N cutoff: the budget may instead depend on the finite target.

Theorem 5.1 (Finite-target sufficiency and exact minimum). For every $K \in \mathbb{N}$, some $N > 0$ satisfies frontier(m, N, K). For $K > 0$, the smallest real-valued sufficient budget is

$$N_{\min}(K) = \frac{1}{g(K-1)}.$$

For $K = 0$, every positive budget works and there is no smallest positive real budget.

Proof. The case $K = 0$ is vacuous. For $K > 0$, positivity gives $g(K-1) > 0$. Antitonicity gives $g(k) \geq g(K-1)$ for every $k < K$, so choosing $N = 1/g(K-1)$ yields $Ng(k) \geq 1$ for all $k < K$. Conversely, every sufficient N must satisfy the constraint at $k = K-1$, hence $N \geq 1/g(K-1)$. □

For an integer candidate count, use $\lceil N_{\min}(K) \rceil$. The highest indexed mode $K-1$ is the binding constraint because it has the smallest coupling among the first K modes.

5.2 Interpretation: Finite Sufficient Budgets in the Model

Theorem 5.1 states that each indexed finite frontier has a finite sufficient budget in the model. It does not guarantee an implementable training process, a valid evaluator, or acquisition of a real-world capability.

This is analogous to the following economic insight: any good can be produced, but not all goods can be produced simultaneously with a fixed budget. The budget constraint creates scarcity, not impossibility.

The paper does not map token counts, compute, or benchmark performance of named language models to the latent budget N or to mode indices. Such a mapping would require a separate measurement model.

---

6. Static Frontier Bounds and Separate Gap Dynamics

6.1 Partial-Sum Consistency Bound

The next inequality is a consequence of the frontier predicate. It constrains admissible counts but does not by itself identify the growth rate of $F_m(N)$.

Theorem 6.1 (Partial-sum frontier inequality). If frontier(m, N, K) holds, then:

$$K \leq N \cdot \sum_{k=0}^{K-1} g(k)$$

Thus every admissible count is bounded by the budget times its partial coupling sum.

Proof. From the frontier condition, $1 \leq N \cdot g(k)$ for each $k < K$. Summing over all $k < K$:

$$K = \sum_{k=0}^{K-1} 1 \leq \sum_{k=0}^{K-1} N \cdot g(k) = N \cdot \sum_{k=0}^{K-1} g(k) \qquad \square$$

If g is summable with $S = \sum_{k\geq 0} g(k)$, applying the inequality to the maximal count gives the weaker general bound

$$F_m(N) \leq NS.$$

This estimate is independent of any recurrence. A pointwise decay law can provide a sharper or exact frontier formula.

6.2 Separate Gap Dynamics

The partial-sum bound controls a static frontier. Conditional on the separate sequence bound, a gap variable describes how the iteration approaches that bound.

Define the gap $G_t = M - K_t$, where M is the explicit sequence bound in Theorem 4.1. This M is not a maximal learnability frontier.

Proposition 6.2 (Gap non-increasing). Under Theorem 4.2's non-regression hypothesis, the gap is non-increasing: $G_{t+1} \leq G_t$.

Proposition 6.3 (Strict gap decrease). Under strict progress — if $K < M$ implies step$(K) \geq K+1$ — the gap decreases by at least 1 per step while below the bound: $K_t + 1 \leq K_{t+1}$ when $K_t < M$.

Corollary 6.4. Under strict progress and the invariant bound, the sequence reaches M in at most $M - K_0$ steps.

Proposition 6.5 (Instant convergence). If $K_0 \leq M$ and step$(K) = M$ for all $K \leq M$, then $K_1 = M$: the bound is reached in a single step.

6.3 Power-Law Coupling: Exact Static Frontier

For a stipulated power-law coupling, the maximal frontier can be computed exactly.

Definition 6.6 (Power-law coupling). For constants $C > 0$ and $\beta > 0$:

$$g(k) = C \cdot (k+1)^{-\beta}$$

No machine formalization of this power-law function, its floor operation, or the inversion argument is claimed.

Theorem 6.7 (Threshold budget for mode k). Under power-law coupling, the real-valued threshold budget for mode k is:

$$N_{\text{thr}}(k) = \frac{(k+1)^{\beta}}{C}.$$

Theorem 6.8 (Exact power-law frontier). For $C > 0$, $\beta > 0$, and $N > 0$,

$$F_m(N) = \left\lfloor (CN)^{1/\beta} \right\rfloor.$$

Proof. Mode k is model-learnable exactly when

$$NC(k+1)^{-\beta} \geq 1 \quad \Longleftrightarrow \quad k+1 \leq (CN)^{1/\beta}.$$

Hence precisely the modes $0, \ldots, F_m(N)-1$ are learnable. If $CN < 1$, the formula gives $F_m(N) = 0$; if $CN = 1$, it gives $F_m(N) = 1$. □

Proposition 6.9 (Steeper coupling, higher threshold). If $\beta_1 < \beta_2$ and $k \geq 1$, then $(k+1)^{\beta_1} < (k+1)^{\beta_2}$. Thus, holding C fixed, the stipulated threshold budget for mode k is larger under β_2.

6.4 A Stylized Exponent Scenario

For illustration only, choose $\beta_{\text{model}} = 3.25$ and define $\alpha_{\text{model}} = 1/\beta_{\text{model}} \approx 0.308$. These are scenario parameters, not estimates fitted here and not the parameter or data exponents reported by Hoffmann et al. (2022). Mapping a language-model loss exponent to the coupling $g(k)$ would require a bridge that this paper does not provide.

Parameter	Value	Implication
α_{model}	$1/3.25 \approx 0.308$	Defined scenario exponent
β_{model}	3.25	Chosen coupling decay rate
Summable ($\beta_{\text{model}} > 1$)?	**Yes**	Finite fixed-N cutoff in the model
$F_m(2N)/F_m(N)$, ignoring floors	$2^{1/3.25} \approx 1.24$	Scenario ratio
$F_m(10N)/F_m(N)$, ignoring floors	$10^{1/3.25} \approx 2.03$	Scenario ratio

Under the chosen scenario, the coupling is summable and the floor-free expression grows by about 24% per budget doubling. For the discrete F_m, the ratio can be undefined when $F_m(N) = 0$ and can remain 1 across budget doublings. This is not a prediction about language-model self-improvement.

The connection to Nagy, *Neural Scaling Laws Formalized: Why Chinchilla Works (A Machine-Verified Derivation)* (2026) remains a proposed modeling bridge. The relationship $\alpha_{\text{model}} = 1/\beta_{\text{model}}$ is a definition inside this scenario, not a derived relationship between the two papers.

6.5 Monotone Threshold Levels

The general possibility of diminishing returns in recursive self-improvement predates this model (Majot and Yampolskiy, 2017). The result below does not establish increasing marginal cost; it only orders the threshold-budget levels associated with successive modes.

Proposition 6.10 (Monotone threshold budgets). The threshold budget $1/g(k)$ associated with mode k is non-decreasing because g is antitone.

Proof. For $k_1 \le k_2$: $g(k_1) \ge g(k_2)$ (antitone), so $1/g(k_1) \le 1/g(k_2)$. □

Proposition 6.11 (Power-law summability). For power-law coupling with $\beta > 1$, $\sum g(k) < \infty$. The exact formula in Theorem 6.8 has sublinear floor-free scale $N^{1/\beta}$; applying this result to natural language requires validating the coupling assumption.

Within the fixed-coupling model, the threshold-budget level is non-decreasing with the mode index. This does not imply that successive budget increments are non-decreasing. Whether either quantity describes practical capability increments or rules out a sudden takeoff requires empirical and algorithmic assumptions outside the formalization.

---

7. Combined Static Threshold Result

7.1 Statement

The combined result packages the static cutoff, finite-target sufficiency, and partial-sum inequality. It contains no recurrence or verifier claim.

Theorem 7.1 (Threshold-model summary). Given a coupling model with summable g:

(i) Fixed N: finite maximal frontier. For every $N > 0$, $F_m(N)$ exists and is finite.

(ii) Target-dependent budget. For every K, some $N > 0$ satisfies frontier(m, N, K); for $K > 0$, the exact minimum is $1/g(K-1)$.

(iii) General summable bound. For $S = \sum_{k \ge 0} g(k)$, $F_m(N) \le NS$.

The following is Lean-style manuscript pseudocode, not generated or checked Lean:

```
theorem self_improvement_limits (m : CouplingModel)
    (h_summable : Summable m.g) :
    ( N :  , 0 < N →  F :  ,  K, frontier m N K  K  F)
    ( K :  ,  N :  , 0 < N  frontier m N K)
    ( N :  , 0 < N →  K :  , frontier m N K →
      (K :  )  N *  k  Finset.range K, m.g k)
```

7.2 Fixed Budget versus Target-Dependent Budget

Part (i) fixes the budget before selecting the maximal frontier. Part (ii) allows the budget to depend on the finite target. These quantifier orders are compatible and say nothing about the trajectory of an iterative system.

Corollary 7.2 (Quantifier contrast). For any coupling model with summable g:

$$\underbrace{\forall N > 0,\ \exists F_m(N) < \infty}_{\text{each fixed budget has a finite maximum}} \qquad \text{and} \qquad \underbrace{\forall K,\ \exists N_K > 0 : \text{frontier}(m, N_K, K)}_{\text{each finite target has a sufficient budget}}.$$

7.3 Interpretation Limits

The result does not resolve claims about explosive improvement or stagnation. The finite-target result supplies a sufficient model budget for each indexed frontier; the fixed-budget result supplies a static maximum. Neither establishes a trajectory.

The model therefore separates two conditional statements: its mode-indexed threshold-budget level is non-decreasing under fixed coupling, while each indexed finite frontier has a sufficient budget under positivity. Neither statement alone establishes practical self-improvement dynamics.

7.4 Positivity Does Not Imply Real-World Reachability

The quantifier contrast depends on positivity inside the model, not on evidence that indexed modes correspond to attainable real capabilities.

Theorem 5.1 needs positivity and antitonicity, not summability. This weak assumption is enough for a finite model budget but not for real-world reachability. The paper does not map FLOPs or named model generations to N, $g(k)$, or mode counts. Such a table would require independently measured coupling and a validated capability-index construction.

---

8. Implications for AI Safety and Governance

8.1 Potential Research Relevance

Inside the model, a fixed positive budget and $g(k) \to 0$ imply a finite maximal frontier, while increasing the budget can move that frontier. This observation may motivate empirical questions about whether any measurable capability ordering has a stable, positive, antitone coupling function that decays to zero.

8.2 Requirements for a Real-World Application

Before the model could inform system assessment or governance, at least four missing components would be needed:

1. an operational capability index with justified ordering and granularity;
2. an estimator for $g(k)$ and evidence that it is stable under algorithmic change;
3. a validated mapping from samples, compute, or verification effort to the scalar budget N;
4. a bridge from the learnability predicate to actual iterative improvement dynamics.

None of these components is supplied here. In particular, N is a sample-budget variable in the mathematics, not automatically a FLOP count or a regulatory compute threshold.

8.3 Verification and Safety

The verification calculation identifies expected accepted yield under a sound-verifier assumption. It does not model false acceptance, correlated candidates, distribution shift, reward hacking, evaluator adaptation, or the effect of accepted examples on future learning. Verification may be useful safety infrastructure, but this paper does not prove that strengthening a verifier raises or lowers the modeled learnability cutoff.

8.4 Explicit Policy Non-Claims

The model does not provide:

- a ceiling certificate for a deployed system;
- a real-world FOOM-risk estimate;
- a safe compute threshold;
- a quantitative basis for export controls or compute caps;
- a claim that verification is universally cheaper than generation;
- necessary or sufficient conditions for AI safety.

The policy relevance is therefore limited to proposing variables and assumptions that an empirical program might try to measure.

---

9. Proof Scope and Claim Map

9.1 Evidence Boundary and Protocol Disposition

The paper's mathematical evidence consists of the conventional arguments printed in Sections 2–7. All numbered statements are paper-only: they have no ssot: annotation, formal_ref, matching proof-kernel declaration, or compiled Lean export. The publication-flow disposition is therefore an explicit **conventional-proof waiver of the G10 statement-level formalization layer**, not an assertion that G10 was satisfied. Ordinary mathematical peer review is required.

The repository also contains a Python-native kernel file with elementary lemmas whose names concern self-improvement, learning theory, optimization, and architecture search. That file is not a formal companion to this manuscript and is excluded from the evidence for the displayed results.

The excluded file does not define the manuscript's CouplingModel, isLearnable, frontier, infinite sums, iteration operator, power-law frontier, or expected-count probability model. Likewise, similarly named declarations do not establish the displayed manuscript theorems. For example:

- bounded_growth proves $C \le K$, $0 \le g$, and $g \le K - C$ imply $C + g \le K$;
- monotone_improvement proves $0 \le g$ implies $C \le C + g$;
- si_scaling_efficiency proves positivity of a ratio under positive numerator and denominator;
- si_power_law_decay proves positivity of C/n under positive inputs.

These are valid local lemmas, but they are not statement-level formalizations of invariant-set boundedness, finite-target sufficiency, the partial-sum inequality, or the exact power-law frontier.

9.2 Paper-to-Proof Claim Map

With the machine-evidence boundary fixed, the following table identifies where each retained mathematical claim is argued.

Manuscript claim	Exact anchor	Evidence actually claimed	Machine formalization
Fixed-budget cutoff and maximal frontier	Theorem 2.11; Definition 2.12	Decay $g(k) \to 0$, antitonicity, and finite-set maximum	None
Bounded recurrence and convergence	Theorems 4.1–4.2	Invariant-set induction; separate non-regression argument	None
Finite-target sufficiency	Theorem 5.1	Positivity, antitonicity, and exact reciprocal budget	None
Partial-sum consistency inequality	Theorem 6.1	Sum the K frontier inequalities	None
Exact power-law frontier	Theorem 6.8	Algebraic equivalence and integer maximum	None
Combined static quantifier statement	Theorem 7.1; Corollary 7.2	Combination of the preceding static manuscript proofs	None

9.3 Reproducibility Status

The paper requires no executable artifact to reproduce the displayed deductions: each follows from the definitions and hypotheses printed at the anchors above. No Lean 4 source has been generated or compiled for these claims. Accordingly, the paper makes no statement-count, kernel-verification, or Lean-verification claim.

---

10. Conclusion: Conditional Frontiers and Open Empirical Gaps

The central result of this work can be stated in one sentence:

> **If an antitone threshold sequence decays to zero, each fixed positive budget has a finite maximal frontier; a separate recurrence is bounded only when it preserves a finite invariant set.**

For any fixed positive budget N, decay gives a finite maximal frontier $F_m(N)$; summability is one sufficient decay condition. For any indexed finite frontier, positivity and antitonicity supply the exact minimum $1/g(K-1)$ when $K > 0$. The partial-sum inequality bounds admissible counts, while pointwise power-law coupling gives the exact frontier. Converting any static statement into a claim about an improvement process requires an explicit bridge to the step function and additional model-to-reality assumptions.

These conditional deductions are argued in the manuscript and are not claimed as machine-formalized. Whether positive decreasing coupling, summability, and any mapping from empirical scaling to g hold for language models remains open. The 24% ratio is arithmetic within a chosen exponent scenario, not an empirical prediction.

10.1 Summary of Results

The model's conclusions and their proof modes can be summarized compactly.

Result	Status	Implication
Fixed-budget cutoff	Manuscript proof	Decay yields a finite maximal frontier $F_m(N)$
Sufficient budgets	Manuscript proof	$N_{\min}(K) = 1/g(K-1)$ for $K > 0$
Partial-sum bound	Manuscript proof	$K \leq N \sum_{k<K} g(k)$; under summability $F_m(N) \leq NS$
Bounded recurrence	Manuscript induction	Preservation gives boundedness; non-regression gives eventual constancy
Power-law scenario	Manuscript algebra	$F_m(N) = \lfloor (CN)^{1/\beta} \rfloor$ exactly
Monotone mode thresholds	Manuscript order argument	The threshold level $1/g(k)$ is non-decreasing; no increasing-increment claim

10.2 Three Implications

For safety. The static model identifies assumptions producing a finite budget frontier, while the separate recurrence lemma identifies an assumed invariant-set condition for bounded dynamics. Neither establishes that these assumptions hold in real systems. The verifier-yield model is motivational and disconnected from both results.

For scaling. The relationship $\alpha_{\text{model}} = 1/\beta_{\text{model}}$ is a definition in the illustrative scenario, not a calibrated relationship to neural scaling exponents. Establishing a connection among loss scaling, covariance spectra, and $g(k)$ is future empirical work.

For policy. No deployed-system frontier certificate follows from this paper. Such a certificate would require all four components in Section 8.2: an operational capability index, a stable estimator for g, a validated budget mapping, and a proved bridge to actual iterative dynamics.

10.3 Limitations and Future Work

Limitations:

1. **Fixed coupling assumption.** The model assumes $g(k)$ is fixed. In practice, algorithmic improvements can change the coupling function. The static frontier applies to a given g; changing g changes the frontier.
2. **Mode independence.** The model treats modes independently. In practice, there may be phase transitions — learning mode k may suddenly unlock modes $k+1$ through $k+10$. Formalizing such transitions in the threshold framework is future work.

3. **Discrete modes.** The model uses $\mathbb{N}$-valued modes. A continuous extension would require new definitions; this paper does not claim that every discrete result transfers unchanged.
4. **No alignment model.** The framework addresses an indexed threshold count, not alignment. A system associated with frontier $F_m(N)$ may or may not be aligned; the model says nothing about what the modes do.

Future work:

1. Determine whether a defensible verifier model can be related to g without assuming away reward hacking.
2. Extend the model to stochastic coupling (random $g(k)$ with known distribution).
3. Prove phase transition conditions: under what coupling structures can mode learning exhibit discontinuous jumps?
4. Test whether measured attention-spectrum quantities can help parameterize coupling decay.

10.4 The Complete Picture

The paper now makes one proof claim only: the displayed results have conventional manuscript proofs, mapped in Section 9. Machine verification, model-to-reality bridges, empirical calibration, oracle mapping, policy conclusions, and Lean export are not claimed.

An optional future program could test whether measured spectral quantities parameterize g. No spectral relation is used in the present proofs, and any connection among a spectral exponent, learning curves, and $F_m(N)$ requires separate validation.

Comparison with Existing Approaches

Approach	Formal?	Machine-checked?	Fixed-budget criterion	Finite-target criterion	Rate bounds
Good (1965); Bostrom (2014)	Conceptual / strategic	No	—	Intelligence-explosion argument	—
Schmidhuber (2003)	Self-referential proof-system model	No artifact claimed here	Utility-improving rewrite criterion	Program-relative	Proof-search cost
Yampolskiy (2015); Majot and Yampolskiy (2017)	Theoretical analyses	No	Computational and convergence limits	RSI taxonomy	Diminishing returns
Anbar Jafari et al. (2025)	Mathematical recursive-improvement model	No artifact claimed here	Conditions and control bounds	Model-dependent	Recursive-growth bounds

Approach	Formal?	Machine-checked?	Fixed-budget criterion	Finite-target criterion	Rate bounds
Davidson et al. (2026)	Semi-endogenous growth model	No	Diminishing-return threshold	Automation-dependent	Explosive-growth condition
Alemohammad et al. (2024)	Empirical recursive-data model	No	Data-distribution degradation	Mitigation by retaining real data	Generational degradation
Kaplan et al. (2020)	Empirical	No	—	Power-law scaling	Fitted exponents
Hoffmann et al. (2022)	Empirical	No	—	Chinchilla scaling	Fitted exponents
This work	**Stylized antitone threshold model**	**No; explicit conventional-proof waiver**	$g(k) \to 0$ + **fixed budget**	**Exact minimum** $1/g(K-1)$	**Partial-sum bound; exact power-law** $F_m(N)$

Scope difference: The closest prior work already studies formal self-modification, recursive-improvement limits, diminishing returns, recursive-data collapse, and explosive-growth conditions. This paper's narrower contribution is the specific antitone-coupling threshold model and its fixed-budget versus target-dependent-budget quantifier contrast. It does not establish that the assumptions describe real systems.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Cursor and its integrated AI models (Claude and GPT families) in order to assist with manuscript drafting, structural editing, and language polishing. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

- Bostrom, N (2014). Superintelligence: Paths, Dangers, Strategies. *Superintelligence: Paths, Dangers, Strategies*.
- Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, 6, 31–88. https://doi.org/10.1016/S0065-2458(08)60418-0.
- Schmidhuber, J. (2003). Gödel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements. *arXiv:cs/0309048*. https://doi.org/10.48550/arXiv.cs/0309048.

- Yampolskiy, R. V. (2015). On the Limits of Recursively Self-Improving AGI. *Artificial General Intelligence*, LNCS 9205, 394–403. https://doi.org/10.1007/978-3-319-21365-1_40.
- Majot, A. M., & Yampolskiy, R. V. (2017). Diminishing Returns and Recursive Self Improving Artificial Intelligence. In *The Technological Singularity: Managing the Journey*, 141–152. https://doi.org/10.1007/978-3-662-54033-6_7.
- Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2024). Self-Consuming Generative Models Go MAD. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2307.01850.
- Anbar Jafari, A., Ozcinar, C., & Anbarjafari, G. (2025). A Mathematical Framework for AI Singularity: Conditions, Bounds, and Control of Recursive Improvement. *arXiv:2511.10668*. https://doi.org/10.48550/arXiv.2511.10668.
- Davidson, T., Halperin, B., Houlden, T., & Korinek, A. (2026). When Does Automating AI Research Produce Explosive Growth? Feedback Loops in Innovation Networks. *NBER Working Paper 35155*. https://doi.org/10.3386/w35155.
- Kaplan, J. et al (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.
- Hoffmann, J. et al (2022). Training Compute-Optimal Large Language Models. *NeurIPS 2022*.
- Christiano, P., et al (2017). Deep reinforcement learning from human preferences. *NeurIPS*.
- Bai, Y., Jones, A., Ndousse, K., et al (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D (2016). Concrete problems in AI safety. *arXiv:1606.06565*.
- Hutter, M (2021). Learning curve theory. *arXiv preprint arXiv:2102.04074*.
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
- Nagy, T. (2026). Generative Portfolio Design: Inverting the Spectral Fenton Representation. *Working paper*.
- Nagy, T. (2026). Neural Scaling Laws Formalized: Why Chinchilla Works (A Machine-Verified Derivation). *Working paper*.
- Irving, G., Christiano, P., & Amodei, D (2018). AI safety via debate. *arXiv:1805.00899*.
- Omohundro, S. M (2008). The basic AI drives. *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. IOS Press.
- Yudkowsky, E (2013). Intelligence explosion microeconomics. *MIRI Technical Report*.