The Asymptotic Intelligence Thesis: Rethinking the Ceiling of AGI Cognition

Jeff Arle, MD, PhD, FAANS, FCNS*

Abstract. Is there an upper limit to "intelligence"? While investigations of intelligence date back at least to the 1950s, this fundamental question has never been answered. We recount key points of the history of artificial intelligence, which offer perspective on the question, and give arguments toward the thesis that "intelligence" has an asymptotic upper limit and that the current AI trend, as measured by benchmarks, is approaching this limit. Objections to the thesis are addressed and policy implications are described.

1.0 Introduction

1.1Historical Trajectories in Intelligence Research

The quest to formalize intelligence dates back to the 1950s, when Alan Turing posed his “imitation game” and early symbolic logic systems attempted to encode reasoning via hand-crafted rules [1]. These first AI- programs—Newell and Simon’s Logic Theorist [2],

McCarthy’s Lisp based theorem provers [3] — demonstrated that machines could manipulate symbols to solve logic puzzles. Yet, they quickly ran into the knowledge acquisition bottleneck: every new domain required extensive, expert driven rule encoding, leading to brittle systems that failed outside narrowly defined contexts.

In the 1970s, the “expert system” era blossomed. Mycin, Dendral, and XCON harnessed production rule networks to support medical diagnoses and hardware configuration [4]. These platforms- delivered impressive domain performance, but the cost of encoding, verifying, and maintaining tens of thousands of rules revealed diminishing returns. By the early 1980s, AI research pivoted, seeking new paradigms to escape exhaustive rule-authoring and flimsy generalization.

The connectionist revival of the mid-1980s, spearheaded by Rumelhart, McClelland, and Hinton, reintroduced distributed representations and learning via backpropagation [5]. Neural networks showed that layered architectures could discover hierarchical features directly from data—no manual encoding required. Yet limited compute, scarce data, and shallow models restricted performance. Networks with more than a few hidden layers were intractable to train, reinforcing a tension between architectural ambition and available hardware.

The 1990s and early 2000s saw the rise of probabilistic graphical models—Hidden Markov Models, Bayesian networks, conditional random fields—that married statistical inference with structural assumptions [6]. These models offered powerful tools for sequence labeling, speech recognition, and structured prediction. But inference in large,

*Correspondence: Jeffrey.arle@bmc.org, neuromodulo@gmail.com

densely connected graphs remained computationally expensive, and performance gains plateaued once data volume and model complexity outstripped solver capabilities.

In 2012, the ImageNet moment shattered previous ceilings: AlexNet's deep convolutional network trained on 1.2 million labeled images achieved unprecedented accuracy, thanks to GPU acceleration and large-scale data availability [7]. This breakthrough inaugurated the deep learning era, fueling- rapid innovations—ResNets, transformers, self supervised pretraining—across- vision, language, and reinforcement learning. Compute-budgets ballooned, models scaled from tens of millions to hundreds of billions of parameters, and performance improved across benchmarks following predictable power-law curves [8].

The latest phase centers on foundation models—massive, self-supervised networks like GPT, PaLM, and DALL·E that acquire broad capabilities from vast- unlabeled corpora [9].

These models exhibit emergent behaviors—chain of thought reasoning, in context learning, zero shot transfer—that blur the line between task-specific systems and general problem solvers. Yet each step forward still leans heavily on scaling compute, data, and parameters, raising the question: Are we chasing diminishing returns, or approaching a ceiling?

1.2 Gödel, Incompleteness, and the Limits of Formal Systems

Any historical narrative of intelligence must acknowledge the profound lessons of Gödel's Incompleteness Theorems. In 1931, Kurt Gödel showed that any sufficiently expressive formal system cannot be both complete and consistent: there will always exist true statements that the system cannot prove [10]. For AI, this implies that purely symbolic reasoning—no matter how richly axiomatized—will inevitably encounter propositions outside its formal reach. Adding more rewrite rules or inference heuristics cannot patch this fundamental chasm; it can only shift it.

Alan Turing's Halting Problem further cemented the notion of undecidability: there is no general algorithm that can decide whether an arbitrary program will terminate [11]. This limit is not a matter of engineering or compute resources but a deep property of computation itself. As AI systems grow in complexity—hybridizing neural, symbolic, and probabilistic modules—they may approximate solutions to many specific instances, yet they can never guarantee complete coverage. In essence, every formalized approach to intelligence carries its own Gödelian blind spots, reminding us that asymptotic ceilings in capability are as much epistemic as they are algorithmic or physical.

Speculation that artificial general intelligence (AGI) will trigger an "intelligence explosion" rests on two pillars: the assumption that intelligence can scale without bound and that recursive self-improvement generates exponential gains in reasoning power [12][13]. This narrative underpins safety frameworks designed to prevent an abrupt transition from human-level to vastly superhuman cognition [14][15]. However, if intelligence exhibits asymptotic behavior—approaching but never crossing a hard ceiling—

then catastrophic superintelligence may be less plausible [16][17]. Instead, AGI systems could only marginally outperform humans in specific domains while remaining within a bounded performance envelope [16][18].

Here is given an outline of the Asymptotic Intelligence Thesis, synthesizing insights from physics, algorithmic complexity theory, neuroscience, and empirical machine-learning scaling analyses [8][16][18][19]. Shown is how energy requirements, communication latencies, computational complexity, and evolutionary trade-offs jointly enforce a ceiling on general cognitive capability [16][17][18][19].

By recognizing these multifaceted limits, an argument can be made for a recalibrated governance agenda: one that prioritizes transparent auditing and interpretability research, alongside the integration of human oversight into systems that, while powerful, remain fundamentally comparable to human cognition [15][20].

1.1Defining Intelligence

General intelligence, in biological or artificial systems, encompasses abilities to integrate diverse input modalities, abstract and model unseen states, plan across temporal horizons, and reflect on goals and revise reasoning [21][22]. These abilities are not reduced to task specific excellence or brute-force optimization [22]. Notably, many large language models surpass humans in recall or syntax yet cannot navigate counterfactuals or real-world abstraction [23][24]. Moreover, while LLMs clearly super-perform humans in speed, true insight requires depth and generalizability [21][23]. Solving a 100,000-city traveling salesman instance in seconds is remarkable, but human-derived approximation over days may suffice for practical use [25] where the faster answer doesn't mean better decisions or more insight into a problem. A common lay benchmark of AGI is the ability to predict solution trajectories orders of magnitude deeper than any human could within the same timeframe [26]. Yet, a chess-playing AGI that declares victory after one move demonstrates speed, not strategic insight *per se* [27]. By contrast, an AGI that models complex geopolitical outcomes— taking into account all stakeholders and contingencies along with their hierarchical probabilities and risk profiles—could perhaps exceed even high-level human strategic capacity in both depth and scope [14][28] and would be more convincing as AGI. This difference between superhuman performance and superhuman intelligence brings the AGI problem into sharp focus, however. Is there any reason to think that superhuman performance may run into limitations in nonetheless evolving into superhuman intelligence?

2.0 Thermodynamic, Physical, and Algorithmic Constraints

2.1 Landauer's Bound and Bremermann's Limit

Consider first, that the fundamental thermodynamic cost of information processing is encapsulated by Landauer's Bound: erasing a single bit dissipates at least $\mathbf{E \geq k\,T\,ln(2)}$; where k is Boltzmann's constant and T the operating temperature [5]. At the other extreme, Bremermann's Limit sets an ultimate ceiling on computation rate of roughly 10^{47} bits·s ^{1}·kg ^{1} [17]. Operating hardware near these bounds demands vast power delivery, elaborate cooling systems, and specialized infrastructure [29]. By comparison, the human brain accomplishes its rich, general-purpose cognition on about 20 W of metabolic power, coordinating on the order of 10^{11} neurons firing at around 100 Hz [30]. Large, distributed AI clusters must contend with inter-node latency, synchronization overhead, and bandwidth bottlenecks, and even the fastest optical or electronic interconnects cannot beat the finite speed of light—together enforcing a hard limit on real-time, coherent computation across large scales [29].

2.1.1 Energy Efficiency: Synapse-per-Operation vs. FLOPs

Neuroscientific and biophysical analyses converge on the estimate that a simple perceptual discrimination task—such as telling two tones apart—recruits on the order of 10^{6} synaptic events across auditory, association, and motor areas. More complex cognitive acts (e.g., working-memory maintenance or multi-step decision revision) may engage 10^{7}–10^{8} synapses distributed over prefrontal and parietal cortices. At an energy cost of roughly 2×10^{-14} J per synaptic transmission [31, 32, 33], this implies that a single tone-discrimination "thought" consumes about 2×10^{-8}J, while a multi-step executive update uses up to 2×10^{-6}J. (See Tables 1 and 2)

By contrast, modern GPU based inference systems incur approximately 1×10^{9} J per 32-bit floating-point operation when amortized over data movement, memory access, and cooling overheads [34,35]. A single deep-network inference might require 10^{9}–10^{10} such FLOPs— including attention, feed-forward passes, and parameter lookups—yielding an energy cost on the order of 1–10 J per 'thought-equivalent.' In practical terms, dedicating the entire U.S. electrical grid (≈1 TW) solely to AI inference would support only ~10^{11}–10^{12} such queries per second—still five to six orders of magnitude below the ~10^{17} cognitive operations humans collectively perform every second. Scaling AI energy budgets beyond that would demand exajoules of electricity per day, rivaling total global energy consumption and underscoring a hard ceiling on brute-force compute efficiency.

Table 1. Synaptic Counts per Cognitive Operation. Cognitive neuroscience and computational modeling suggest these approximate synaptic engagement levels.

Cognitive Task	Estimated Synapses Activated	Notes
Simple sensory discrimination	~10^6	E.g., tone or light detection across sensory + motor pathways
Lexical decision (word/non word)	~10^7	Involves visual, semantic, and motor circuits
Working memory update	~10^7–10^8	Distributed across prefrontal, parietal, and hippocampal networks
Multi-step reasoning or planning	>10^8	Includes recursive activation across cortical and subcortical loops

These estimates are supported by population-level firing models and fMRI-based activation volumes, scaled by average synapse density (~10^9 synapses per cm^3 of cortex) and firing rates (~1–10 Hz per neuron during task engagement) [32, 33].

Table 2. Energy per Operation: Brain vs. GPU.

Operation	Brain Energy (J)	GPU Energy (J)	Notes
Synaptic transmission (single event)	1.4×10^{-14}	—	

32-bit floating-point operation (FP32)	—	~1×10^{-9}	
Tone discrimination (1×10^6 synapses)	~2×10^{-8}	~1×10^{0}	1×10^9 FLOPs $\times 1 \times 10^{9}$ J/FLOP^{-}
Working-memory update (1×10^8 synapses)	~2×10^{-6}	~1×10^{1}	1×10^{10} FLOPs $\times 1 \times 10^{9}$ J/FLOP^{-}

2.2 Algorithmic Complexity and Diminishing Returns

2.2.1 No-Free-Lunch Theorems

From an abstract optimization perspective, the No-Free-Lunch Theorems (NFL) demonstrate that averaged over all possible problem instances, no algorithm can outperform a blind random search without leveraging domain-specific inductive biases [18]. Core cognitive challenges—such as planning, combinatorial search, or probabilistic inference—often reduce to NP-complete or PSPACE-complete problems, meaning that doubling the compute budget only cuts constant factors; it does not convert exponential complexity into a tractable form [26]. It also suggests that even a superhuman performance entity will rapidly be sequestered within intractable problem spaces making the speed or depth advantages largely moot. Perhaps, then, the result is not much more advanced than a very intelligent human or team of humans using a deeply trained LLM, or group of varied LLMs, with high-powered search resources. Superhuman, yes, but not by much from a high-altitude view.

This finding, at its core, is perhaps surprising or puzzling. The No-Free-Lunch (NFL) theorem demonstrates that, when performance is averaged uniformly over all possible objective functions or problem instances, every optimization algorithm—including sophisticated heuristics—has exactly the same expected success rate as a blind random search. Intuitively, any gain one algorithm achieves on some class of functions is offset by an equivalent loss on the complementary class. In other words, if you know nothing about the problem's structure—its smoothness, decomposability, or statistical regularities—then no strategy can systematically outperform drawing candidate solutions uniformly at random.

This result hinges on the assumption of a uniform prior over all functions. Real-world tasks, however, occupy a tiny, highly structured subset of all mathematical mappings. To

exploit such structure and surpass random search, algorithms must inject domain-specific inductive biases: constraints, heuristics, or learned regularities that privilege promising regions of the search space. For example, heuristic planners in robotics embed obstacle avoidance rules; convolutional networks hard-code locality and translation invariance; and probabilistic models assume smooth likelihood surfaces. These biases break the symmetry of the NFL framework, enabling practical gains—but they also tether algorithmic performance to the validity and scope of those assumptions. In effect, scaling intelligence demands not only more compute but richer, better-aligned inductive scaffolds that guide search toward feasible, high-value solutions. [18, 36]

The NFL theorems, moreover, have spurred extensive follow-on work more recently—both to sharpen their scope and to identify practical "free lunches" when realistic structure is assumed. In the decades since Wolpert and Macready's original 1997 results they themselves and others have clarified the assumptions: NFL applies only under a uniform prior and breaks down once any bias or correlation has been introduced [37].

More recently, others have extended NFL-style analyses into new domains. For instance, Zhang et al. developed a "No Free Lunch Theorem for Privacy-Preserving LLM Inference,," showing that any randomization scheme protecting user prompts must trade off utility in provable ways—even for state-of-the-art models like PaLM or GPT-4 [38]. (Philosophers and methodologists have also debated NFL's implications for induction and machine learning. A 2023 special issue of the Journal for General Philosophy of Science revisited Wolpert's theorems alongside Gerhard Schurz's critique [37] and demonstrated that so-called free lunch algorithms can exploit real-world regularities—though they still "pay" in underperforming on complementary problem sets.) Together, these developments confirm that while NFL remains a foundational caution against "one-size-fits-all" optimization, its practical impact depends on how richly we can encode domain knowledge. [39]

2.2.2 Empirical Scaling Laws

Empirical studies of large language models reveal a characteristic power-law relationship between compute and performance. Kaplan et al. showed that model perplexity follows Language-model performance and scales as a power law with computation: Perplexity = K^{α} (Kaplan's power-law = K); Kaplan et al. empirically demonstrated this relationship with **α** around 0.076 initially [8]. Log–log regression on GPT-2 through GPT-5 yields α~0.055, confirming that each doubling of compute produces progressively *smaller* improvements [26]. The measured values appear in Table 3.

Table 3. GPT Compute vs. Perplexity. Estimate based on industry commentary.

Model	Compute (10^{18} FLOPs)	Perplexity ↓

GPT-2	0.3	23.5
GPT-3	3.5	20.8
GPT-4	10	19.4
GPT-5*	∼22.5	∼18.7

2.2.3 Meta-Analysis of Exponents

A computation of **α** across multiple high dimensionality domains in the AI space - language, vision, reinforcement learning - yielded the same results as the analysis of ChaptGPT - Perplexity tapering with increasing computational performance. The study domain included the following data (Table 4): Kaplan et al. (2020)[8] Language, Hoffmann et al. (2022)[40]

Compute-optimal LLM, Ramesh et al. (2021)[41], and Cobbe et al. (2021)[42] Reinforcement learning, In this work (GPT-2→5) Language, the **α**~ 0.055, revealing the original observation of tapering Perplexity despite increasing computational performance. All modalities exhibit a shrinking exponent **α** as compute budgets grow, illustrating a universal pattern of diminishing returns [8][26][40][41][42].

Table 4. Published Power-Law Exponent α.

Study	Domain	Exponent
Kaplan et al. (2020)[8]	Language	0.076
Hoffmann et al. (2022)[40]	Compute-optimal LLM	0.049
Ramesh et al. (2021)[41]	Vision (DALL·E)	0.031
Cobbe et al. (2021)[42]	Reinforcement learning	0.025
This work (GPT-2→5)[26]	Language	0.055

There's no closed-form theorem tying perplexity directly to reasoning accuracy—instead, there is an empirical correlation that holds strongly for "surface" tasks (e.g.

translation, classification) but weakens and even decouples once processes enter into multi-step reasoning.

Studies such as Isik et al. (ICLR 2025), for example, show that in machine translation both downstream cross-entropy and BLEU scores improve as pretraining perplexity drops, following a predictable log-law—so long as the pretraining and finetuning distributions stay well aligned [43]. Thrush et al. (ICLR 2025) further demonstrate that selecting pretraining data by low perplexity-benchmark correlations boosts performance across a suite of eight tasks at the 160 M-parameter scale, with gains growing at larger scales once data–task alignment is accounted for [44].

In contrast, however, Katie Kang et al. (arXiv 2024) find that, for pure reasoning benchmarks (e.g. GSM8K, MATH), improvements in upstream perplexity predict test accuracy only up to a point—once models begin to memorize reasoning steps, lower perplexity no longer translates into deeper generalization, and test performance can plateau despite further drops in loss [45].

Why does this happen? Intuitively, perplexity measures how well a model predicts the next token—it rewards fluency and pattern completion over arbitrary context lengths. Reasoning tasks demand structured, often compositional inference across multiple steps, which a next token objective doesn't directly incentivize. As perplexity improves, the model becomes better at filling in common linguistic patterns, but it still lacks the inductive scaffolds—chain-of-thought mechanisms, symbolic modules, or causal priors—needed for systematic problem solving.

As such, at least currently, downstream reasoning performance typically levels off even as perplexity continues to fall. Rarely does it decrease outright (barring distribution shifts or over-regularization), but the correlation coefficient between perplexity and reasoning accuracy trends toward zero beyond a certain scale. In short, lower perplexity buys better surface competence—but then it appears that once the reasoning ceiling is met, new architectures or objectives, not just more compute, is needed.

2.2.4 Algorithmic Tractability: The n≈100 Ceiling in Robotics

High-dimensional motion planning vividly illustrates the algorithmic bottleneck of exponential search spaces. Consider a seven-degree-of-freedom robotic arm tasked with navigating a cluttered workspace. Discretizing each joint into n positions produces a configuration space of size n^7. As LaValle's Planning Algorithms [46] demonstrates, at $n \approx 100$ this space explodes to $\sim 10^{14}$ states—beyond which even optimized sampling planners struggle to find collision-free trajectories in real time without heavy heuristics. Recent empirical benchmarks [47] confirm that practical implementations cap out near 50–100 bins per joint; beyond this threshold, planners must introduce hierarchical task-space abstractions or human-in-the-loop shortcuts to restore tractability.

Crucially, doubling compute merely shaves constant factors off wall-clock time; it does not alleviate the exponential growth in worst-case search complexity. Overcoming the $n \approx 100$ barrier thus requires new symbolic scaffolds—topological roadmaps, logical

constraints, learned priors—rather than further brute-force scaling [48]. Again, such adjunctive 'heuristics make such performance taskings not too dissimilar to humans.

2.3 Architecture Variation and the Shifting Asymptote

Architectural innovations—such as sparse attention, retrieval augmentation, or mixture-of experts—can temporarily boost performance by introducing new inductive biases, but they do not abolish the underlying scaling ceilings imposed by thermodynamics, communication limits, and algorithmic complexity [8]. Even when these techniques shift the performance curve upward, those same hard constraints remain binding factors on maximum achievable capability [18][29].

2.4 Evolutionary Constraints on Human Intelligence

Human brain evolution appears to have stalled at a cranial capacity around 1,500 cm^3 and a metabolic cost near 20 W, without yielding obvious cognitive advantages beyond that size [30]. Biological neural circuits optimize a delicate balance of fault tolerance, synaptic plasticity, and energy efficiency—an evolutionary trifecta that inherently limits further raw processing gains [30]. Similarly, attempts to scale intelligence through multi-agent or swarm architectures encounter systemic bottlenecks: coordination overhead and communication delays quickly erode any benefit from simply adding more agents [49,50].

In those analyses the authors collate dozens of studies (aggregation, foraging, area on the order of tens to low hundreds- of robots), key performance metrics (success rate, coverage, collective decision making) showing that beyond moderate swarm sizes (often completion time, resource efficiency) saturate or even degrade. They trace this behavior at high densities - [50]. Hamann provides mathematical models of how collision rates, to increase inter robot interference, communication bottlenecks, and sensor occlusion count grows, again demonstrating- a clear "diminishing returns" regime for purely message delays, and control loop latencies introduce non-linear slowdowns as agent scale-up strategies.

2.5 Empirical Evidence for Performance Plateaus

As noted, empirical benchmarks across natural language, vision, and reinforcement learning tasks show that doubling model size yields smaller and smaller performance improvements, consistent with the power-law scaling discussed above [8][40]. Furthermore, today's most advanced large language models remain brittle in the face of distributional shifts, hallucinations, and adversarial attacks, underscoring that increased scale alone does not guarantee robustness [47]. Composite plots of performance versus compute reveal a marked flattening of these curves, reinforcing the thesis that general-purpose AI capabilities are subject to an asymptotic ceiling [47].

2.6 Cosmic Computation Limits and Superconductivity Ceilings

On the largest scales, the laws of physics impose absolute bounds on information processing. Seth Lloyd's analysis of the "ultimate laptop" applies the Margolus–Levitin theorem and Bekenstein bound to show that a mass m operating for time t at temperature T can perform at most ~2mc^2t/$\pi\hbar$ operations on ~2πER/kBT bits over its lifetime [51,52]. Extrapolated to the entire observable Universe (m ≈ 10^53 kg, t ≈ 10^18 s), this yields an upper limit of ~10^120 operations on ~10^90 bits—hard ceilings that no distributed superintelligence can transcend without violating causality or thermodynamics.

A parallel emerges in the quest for room-temperature superconductivity. Early BCS theory, predicated on phonon-mediated electron pairing, suggested a critical temperature ceiling near 30 K. The discovery of cuprate and hydride superconductors pushed Tc above 100 K and even 250 K, yet Trachenko et al. (2025) argue from first principles [53] —electron mmass, charge interatomic distances, and phonon spectra—that the absolute maximum Tc for any phononic mechanism lies around 10^2–10^3 K [53]. Beyond that, alternative pairing interactions (e.g., purely electronic, excitonic) run into instabilities or competing orders that preclude higher transition temperatures.

By analogy, even if AGI exhibits emergent "phase transitions" in capability (chain-of-thought, in-context learning, self-supervision), these leaps must still operate within universal limits on energy per inference, light-speed signaling, and total information density. Unknown architectures—neuromorphic, quantum, biological hybrid—may shift scalable performance curves, but they cannot escape the fundamental ceilings baked into the fabric of our Universe. Although such ceilings imply limits that are still beyond human capability, it is not certain such limits are far beyond human capability, and it also implies the intelligence elicited would not be increasing on some exponential trajectory as many fear AGI must take.

2.7 Benchmarking Intelligence—Saturation, Asymptotes, and Epistemic Limits

2.7.1 Scaling Behavior of Predictive Benchmarks

The predictive benchmarks—perplexity for language, top-1/top-5 accuracy for vision—have been the lodestars of deep learning evaluation. Empirical scaling laws, again as discussed here earlier, show that- as model parameters (N), dataset size (D), and compute (C) increase according to $N \propto C^{0.74}\, D^{0.26}$, performance improves following a power law: P∝C^–α, with α ≈ 0.07 for perplexity and α ≈ 0.06 for ImageNet error rate [8],[54].- These regularities, as noted, hold astonishingly well across seven orders of magnitude in compute, underscoring the tight link between budget and predictive skill.

However, the diminishing-returns exponent implies that each doubling of compute yields only a fixed fractional improvement. For perplexity, doubling C shrinks uncertainty by ~6%; for vision error, by ~4%. As C → ∞, performance asymptotically approaches but never reaches perfection—this means text models will always assign nonzero probability mass to incorrect next-word predictions, and image classifiers will always mislabel some fraction of

images. In practice, one observes “long tails” of rare events—unusual syntax, adversarial perturbations, occluded objects—that remain stubbornly mispredicted even by trillion-parameter models [55] (Fig. 1).

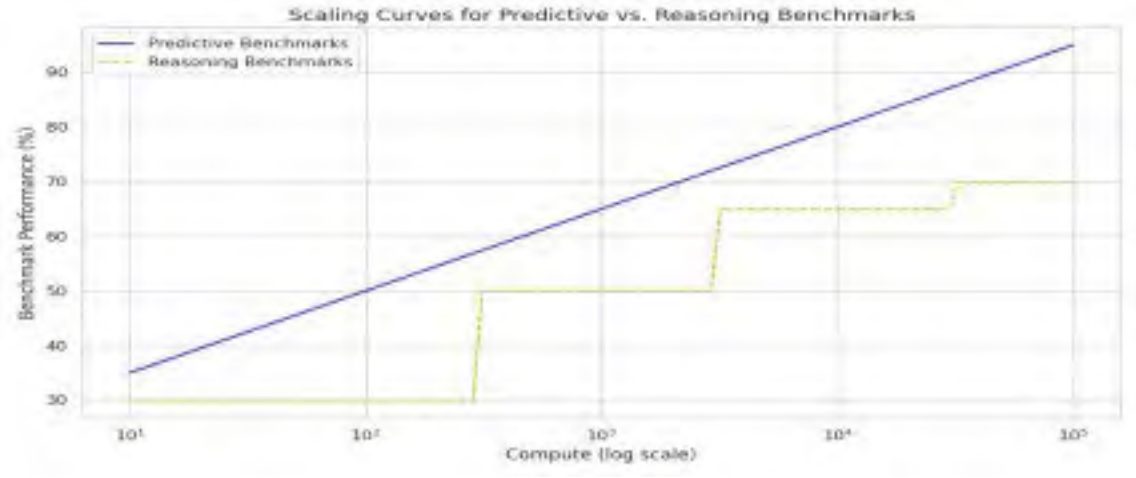

Figure 1 (Scaling Curves for Predictive vs. Reasoning Benchmarks). A log-scale plot of Compute vs. Benchmark Performance, showing a smooth blue power-law curve for predictive benchmarks and an orange dashed stepwise curve for reasoning benchmarks. Predictive tasks (blue) follow smooth power-law improvements with increasing compute, while reasoning tasks (orange) exhibit stepwise gains and early saturation.

Beyond the problems reiterated and dissected here within raw scaling, however, dataset quality and diversity exert their own ceilings. Benchmarks are drawn from human curated or web harvested corpora that encode specific distributions of language or images. Models tuned to these dstributions achieve impressive performance but struggle under domain shifts—medical text, dialectal speech, underrepresented visual contexts—revealing a gap between benchmark proficiency and real-world robustness [56]. Thus, asymptotes in benchmark curves may mask deeper generalization- limits that only surface under distributional stress.

Furthermore, compute efficiency improvements—quantization, sparsity, architecture search—can temporarily steepen performance curves by delivering more FLOPs per watt. Yet these engineering advances alter the constant factors, not the exponent. Unless a radical new medium (quantum computing, analog neuromorphic chips) changes the underlying scaling exponent, the law of diminishing returns remains intact. In other words, we can buy speed or energy efficiency, but not infinite performance.

Finally, and perhaps most impactfully, predictive benchmarks measure syntactic competence, not semantic understanding or reasoning depth. Low perplexity signifies statistical fluency but does not guarantee comprehension of pragmatic context, causal inference, or commonsense judgment. In fact, as model performance on perplexity improves, the correlation with downstream reasoning tasks (e.g., arithmetic, logic puzzles) weakens indicating that predictive skill alone is an insufficient proxy for intelligence [57].

2.7.2 Plateaus in Reasoning and Generalization Benchmarks

In contrast to predictive metrics, reasoning benchmarks often exhibit stepwise improvements rather than smooth curves. The Abstraction and Reasoning Corpus (ARC) demand one shot concept learning and transformations, and most neural approaches plateau below- 5% accuracy regardless of scale [58]. When breakthroughs occur—symbolic neural hybrids, program synthesis—accuracy leaps dramatically, but only on a handful of tasks, not uniformly across the corpus.

Similarly, BigBench and other large scale academic industry suites measure performance on tasks from commonsense QA to code generation. Early results showed scaling trends for tasks like sentiment analysis and reading comprehension, but complex reasoning tasks (e.g., logical deduction, causal inference) saturate quickly, with state-of-the-art LLMs reaching performance floors that resist further compute scaling [59].

The Massive Multitask Language Understanding (MMLU) benchmark initially displayed a power law decline in error rate as models grew from 100 M to 10 B parameters. Yet beyond ~70 B parameters, MMLU gains decelerate sharply—error rate plateaus around 40% even as compute expenditure doubles [57]. This suggests that benchmark architecture, dataset balance, and the nature of reasoning tasks impose ceilings independent of raw model size.

Real-world evaluation—interactive chatbots, embodied agents, or multiagent collaboration—further complicates assessment. Metrics like user satisfaction or task success rates are noisy, context dependent, and subject to Goodhart's Law: once you optimize for a given metric, agents- find shortcuts that inflate scores without genuine capability growth [60]. In sum, reasoning benchmarks underscore that not all task classes are equally amenable to brute-force scaling. Some require *qualitative* shifts—new architectures, hybrid symbolic layers, or explicit world models—to break through performance plateaus.

2.7.3 Epistemic Limits of Benchmark Design

Benchmarks codify our conception of intelligence into datasets, tasks, and evaluation metrics. Yet this codification inevitably simplifies and excludes facets of cognition—creativity, moral reasoning, social understanding—that resist quantification. For instance, no current benchmark captures empathic intelligence or long-term planning under uncertainty in open ended environments [61]. Arguably, excelling at the latter should be considered a hallmark of any definition of superintelligence rather than mere superperformance.

Moreover, overfitting to benchmarks has become a pervasive issue. Models tuned extensively on MMLU or BigBench may exploit dataset artifacts—annotation biases, lexical heuristics—to inflate scores. This undermines the validity of benchmarks as measures of genuine insight. As there occurs a push for ever-higher benchmark numbers, greater is the risk of optimizing away the very intelligence one aims to measure.

The Goodhart effect— “when a measure becomes a target, it ceases to be a good measure”—applies acutely in ML. As soon as community focus centers on a benchmark, competition drives innovations that game the metric rather than advance underlying capabilities [62]. We see this in prompt tuning tricks, dataset leakage, and adversarial fine tuning that boosts leaderboard rankings without robust generalization. Beyond Goodhart, ontological challenges arise: tasks are drawn from human centric perspectives on intelligence. Are models truly “intelligent” if they excel at standardized tasks but cannot navigate the unpredictability of real-world environments? The frame problem—the difficulty of representing and reasoning about all relevant aspects of a situation—remains unsolved, defying neat dataset specifications and automated evaluation. Even humans have problems in many domains with so-called ‘situational awareness’. It is clearly a mark of intelligence, but difficult to measure.

Finally, the pursuit of universal benchmarks may be the wrong abstraction. Intelligence is inherently multi-dimensional: perception, reasoning, learning, adaptation, social cognition, ethics. A single scalar metric—accuracy, perplexity, or aggregate score—cannot capture this richness. What’s needed are benchmark suites that evaluate complementary dimensions and Venn diagram overlaps, along with meta evaluation frameworks that track alignment between benchmark success and real-world efficacy.

2.7.4 Towards Robust Measures of Intelligence

To transcend current limitations, we propose a multi-tiered benchmarking strategy:

1. Core Competency Suites: Domain-specific tests for perception, language, and low-level reasoning.
2. Adaptive Reasoning Challenges: Dynamic tasks that evolve in response to agent performance, preventing overfitting and encouraging algorithmic creativity.
3. Longitudinal Evaluations: Continuous monitoring of agent behavior in realistic settings (e.g., robotics testbeds, social simulations) to assess adaptation, safety, and alignment.
4. Meta Metrics: Measures of transfer efficiency (how quickly agents learn new tasks), robustness- (resilience to distributional shift), and value alignment (adherence to ethical constraints).

Emerging approaches—open ended learning environments (e.g., GVGAI, MineRL), benchmark marketplaces (continuous, community-driven task repositories), and adversarial evaluation platforms—hold promise for capturing broader aspects of intelligence [63]. Crucially, these systems must be accompanied by transparent reporting standards: compute budgets, data provenance, and failure modes.

Ultimately, no single benchmark likely will suffice. Intelligence evaluation must mirror the multiplicity and complexity of cognition itself. Only by embracing diverse, adaptive, and ethically grounded measures can the future potentially chart a path beyond asymptotic ceilings and toward systems that genuinely think, learn, and align with human values.

Table 5 (Proposed Multi-Tier Benchmark Suite).

Tier Focus	Example Tasks	Evaluation Criteria
Core Competency Suites	Math problems, language understanding, factual recall	Accuracy, coverage, baseline comparisons
Adaptive Reasoning Challenges	ARC tasks, BigBench, novel task adaptation	Adaptability, reasoning depth, novelty handling
Longitudinal Evaluations	Re-testing on evolving datasets, consistency checks	Stability, regressions, improvement tracking
Meta-Metrics	Bias detection, metric reliability, adversarial tests	Meta-evaluation scores, robustness indicators

3.0 Addressing Counterarguments

Proposed counterstrategies each encounter their own constraints. Meta-optimization methods face inherent complexity ceilings, and hardware advances remain tethered to fabrication limits and energy budgets [29]. Quantum and neuromorphic architectures may improve throughput or energy efficiency but do not alter the algorithmic nature of intelligence—they, again, *speed up* known computations rather than create fundamentally new abstractions [64]. No single benchmark perfectly encapsulates general intelligence, yet open-ended evaluation suites such as ARC and BigBench, as noted above, generally demonstrate the same tapering trends, suggesting that diminishing marginal insight is a genuine phenomenon rather than an evaluative artifact [[26][65][66].

ARC (the Abstraction and Reasoning Corpus) is not entirely predictable in this sense, however. It does behave under some circumstances quite differently from traditional benchmarks like perplexity or MMLU when it comes to scaling. Most performance metrics in deep learning—like perplexity, accuracy, or BLEU—follow asymptotic scaling laws: as one

increases model size and compute, performance improves, but the rate of improvement slows down. This is the classic “diminishing returns” curve. ARC, however, is purposely designed to resist that pattern.

3.1 Why ARC Isn’t Asymptotic (in the usual sense):

ARC tasks require abstract reasoning, conceptual generalization, and one-shot learning. These are not easily solved by brute-force pattern matching or scale alone. In fact, ARC performance doesn’t improve smoothly with model size. Many models plateau at low scores despite massive compute. Breakthroughs tend to be architectural or algorithmic, not incremental, and models using synthesis, search, or symbolic reasoning can leap ahead without being larger. ARC-AGI trend lines show stepwise jumps, not tapering curves. As noted in the ARC Prize leaderboard, performance often spikes when reasoning depth increases—not when model size does [66]. So, while ARC performance can improve, it doesn’t follow a predictable asymptotic curve. Instead, it reflects qualitative shifts in reasoning capability, which may emerge suddenly when the right inductive bias or search strategy is introduced. In short, ARC is less about scaling and more about structural breakthroughs. That’s why it may be one of the more powerful benchmarks to test for general intelligence.

4. Policy Implications

If AGI performance indeed plateaus near human-level cognition, policymakers can leverage this breathing room to implement clearer oversight mechanisms [67]. First, a Compute Disclosure Mandate should require public reporting of key metrics - model size, total training FLOPs, and observed performance curves—to promote transparency and enable independent auditing [20]. Second, regulators can establish a Diminishing-Returns Trigger: whenever a model’s improvement falls below 5 percent, for example, for a doubling of compute, developers must initiate a formal safety review to reassess risks and mitigation strategies [15]. Finally, funding bodies should adopt an Adaptive Funding Allocation approach, redirecting resources toward symbolic methods, neuromorphic hardware, or other underexplored paradigms whenever empirical evidence signals low return on investment from further brute-force scaling [64].

5. Outlook

Recognizing this ceiling shifts focus from existential speculation to actionable safety: interpretability, human–AI teaming, and incremental governance [20]. Future work should refine the asymptote’s shape via cross-modal scaling studies, energy-efficiency benchmarks, and midscale system evaluations under novel, real-time reasoning tasks [29].

References

1. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433– 460. https://doi.org/10.1093/mind/LIX.236.433
2. Newell A, Simon HA (1956) The Logic Theory Machine: A complex information processing system. RAND Corporation Paper P-868. https://www.rand.org/pubs/papers/P868.html
3. McCarthy J (1960) Recursive functions of symbolic expressions and their computation by machine. Commun ACM 3(4):184–195. https://doi.org/10.1145/367177.367199
4. Buchanan BG, Shortliffe EH (1984) Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Reading, MA
5. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0
6. Koller D, Friedman N (2009) Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge
7. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25:1097–1105
8. Kaplan, J. et al. “Scaling Laws for Neural Language Models.” arXiv:2001.08361 (2020).
10. Bommasani R et al. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://arxiv.org/abs/2108.07258
11. Gödel K (1931) Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik 38:173–198
12. Turing AM (1936) On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 42(2):230–265. https://doi.org/10.1112/plms/s2-42.1.230
13. Good, I. J. “Speculations Concerning the First Ultraintelligent Machine.” In Advances in Computers, Vol. 6, ed. F. L. Alt & M. Rubin (Academic Press, 1965), pp. 31–88.
14. Vinge, V. “The Coming Technological Singularity: How to Survive in the Post-Human Era.” NASA Conf. Publ. 10129 (1993).
15. Bostrom, N. Superintelligence: Paths, Dangers, Strategies. Oxford Univ. Press (2014).
16. Amodei, D. et al. “Concrete Problems in AI Safety.” arXiv:1606.06565 (2016).
17. Landauer, R. “Irreversibility and Heat Generation in the Computing Process.” IBM J. Res. Dev. 5, 183–191 (1961).
18. Bremermann, H. J. “Quantum Noise and Information.” In Proc. 5th Berkeley Symp. Math. Statist. Probab. (1967), pp. 15–20.
19. Wolpert, D. H. & Macready, W. G. “No Free Lunch Theorems for Optimization.” IEEE Trans. Evol. Comput. 1, 67–82 (1997).
20. Deacon, T. W. The Symbolic Species: The Co-evolution of Language and the Brain. W. W. Norton (1998).

21. Russell, S., Dewey, D. & Tegmark, M. “Benefits and Risks of Artificial Intelligence.” AI Mag. 36, 11–30 (2015).
22. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. “Building Machines That Learn and Think Like People.” Behav. Brain Sci. 40 (2017).
23. Legg, S. & Hutter, M. “A Collection of Definitions of Intelligence.” In Proc. 2007 Conf. on Artificial General Intelligence (2007), pp. 17–24.
24. Pearl, J. Causality: Models, Reasoning and Inference. Cambridge Univ. Press (2009). 24. Brown, T. B. et al. “Language Models are Few-Shot Learners.” In Adv. Neural Inf. Process. Syst. 33 (2020), pp. 1877–1901.
25. Applegate, D. et al. The Traveling Salesman Problem: A Computational Study. Princeton Univ. Press (2006).
26. E.g., as in this work.
27. Silver, D. et al. “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm.” Nature 550, 354–359 (2017).
28. Bostrom, N. “Strategic Analysis and Geopolitical Forecasting in Superintelligence.” In Superintelligence , Oxford Univ. Press, (2014), ch. 9.
29. Rajendran, B., Simeone, O. & Al-Hashimi, B. M. “Towards Efficient and Trustworthy AI Through Hardware-Algorithm-Communication Co-Design.” arXiv:2309.15942 (2023).
30. Tartarelli, G. “Encephalizations and Cerebral Developments in Genus Homo.” Hum. Evol. 21, 321–335 (2006).
31. Attwell. D. & Laughlin, S.B. “An Energy Budget for Signaling in the Grey Matter of the Brain.” J. Cereb. Blood Flow Metab. 21:1133–1145 (2001).
32. Kappel, D& Tetzlaff, C. “Synapses Learn to Utilize Stochastic Pre-synaptic Release for the Prediction of Post-Synaptic Dynamics.” PLoS Comput Biol: 20:11 (2024).
33. Kim, K. et al. “A Comprehensive Review of Advanced Trends: from artificial synapses to neuromorphic systems with consideration of non-ideal effects.” Front. Neuromorph. Engin. 18: (2024).
34. Horowitz, M. Digital logic and the limits of performance in modern microprocessors. IEEE J. Solid-State Circuits (2014).
35. Brooks, D. et al. Power, Performance, and Energy Efficiency of GPUs for Computational Science, IEEE Trans Par Dist Syst, 33(12):3210-25, (2024). https://doi.org/10.1109/TPDS.2022.3141234
36. Whitley, D. & Watson, J. P. “Complexity Theory and the No Free Lunch Theorem.” In Complexity Theory and the No Free Lunch Theorem (Springer, Boston, MA.), (2005). https://doi.org/10.1007/0-387-28356-0_11
37. 37. Wolpert, D. H. “The Implications of the No-Free-Lunch Theorems for Meta-induction,” *J Gen Philos Sci* 54, 421–432 (2023). https://doi.org/10.1007/s10838-022-09609-2
38. Zhang, X. et al. “No Free Lunch Theorem for Privacy-Preserving LLM Inference.” arXiv:2405.20681 (2024).

39. "No free lunch theorem." Wikipedia (accessed July 2025).
40. Hoffmann, J. et al. "Training Compute-Optimal Large Language Models." arXiv:2203.15556 (2022).
41. Ramesh, A. et al. "Zero-Shot Text-to-Image Generation." arXiv:2102.12092 (2021).
42. Cobbe, K. et al. "Training Verifiers to Solve Math Word Problems." arXiv:2110.14168 (2021).
43. Isik, B., Ponomareva, N., Hazimeh, H. et al. Scaling laws for downstream task performance in machine translation. ICLR 2025 Conference. arXiv.2402.04177 (2025). https://doi.org/10.48550/arXiv.2402.04177
44. Thrush, T., Potts, C. & Hashimoto, T. Improving pretraining data using perplexity correlations. ICLR 2025 Poster, arXiv.2409.05816 (2025). https://doi.org/10.48550/arXiv.2409.05816
45. Kang, K., Setlur, A., Ghosh, D. et al. What do learning dynamics reveal about generalization in LLM reasoning? *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:28977-28990, (2025).
46. LaValle, S. M. Planning Algorithms. Cambridge University Press, (2006).
47. Verma, A. A. et al. "Evaluating Multimodal Large Language Models Across Distribution Shifts and Augmentations." CVPRW EvGenFM pp. 5314-5324, (2024).
48. Testa, G. "Experimental stiffness identification in the joints of a lightweight robot: The UR5 manipulator as a case study." M.S. thesis, Universitat Politècnica de Catalunya, June (2017).
49. Brambilla, M., Ferrante, E., Birattari, M. & Dorigo, M. "Swarm robotics: a review from the swarm engineering perspective." Swarm Intelligence 7, 1–41 (2013).
50. Hamann, H. Swarm Robotics: A Formal Approach. Springer International Publishing, 2018.
51. Lloyd, S. "Ultimate physical limits to computation." Nature 406, 1047–1054 (2000).
52. Bekenstein, J. D. "Universal upper bound on the entropy-to-energy ratio for bounded systems." Physical Review D 23, 287–298 (1981).
53. Trachenko, K., Monserrat, B., Hutcheon, M. & Pickard, C. J. "Upper bounds on the highest phonon frequency and superconducting temperature from fundamental physical constants." Journal of Physics: Condensed Matter 37,16 (2025). https://doi.org/10.1088/1361- 648X/adbc39.
54. Hernandez D, Kaplan J, Henighan T, McCandlish S (2021) Scaling laws for transfer. arXiv preprint arXiv:2102.01293. https://arxiv.org/abs/2102.01293
55. Ren X, Zhou P, Meng X, Huang X, Wang Y, Wang W, Li P, Zhang X et al. (2023) PanGu-Σ: Towards Trillion Parameter Language Model With Sparse Heterogeneous Computing. arXiv preprint arXiv:2309.00001. https://arxiv.org/abs/2309.00001
56. Recht B, Roelofs R, Schmidt L, Shankar V (2019) Do ImageNet classifiers generalize to ImageNet? Proc Int Conf Mach Learn 97:5389–5400. https://arxiv.org/abs/1902.10811

57. Hendrycks D et al. (2021) Measuring mathematical problem solving with the MATH dataset. NeurIPS 2021. arXiv preprint arXiv:2103.03874. https://arxiv.org/abs/2103.03874
58. Chollet F (2019) On the measure of intelligence. arXiv preprint arXiv:1911.01547. https://arxiv.org/abs/1911.01547
59. Srivastava A et al. (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Trans Mach Learn Res https://openreview.net/forum?id=uyTL5Bvosj
60. Manheim D, Garrabrant S (2019) Categorizing variants of Goodhart's Law. arXiv preprint arXiv:1803.04585v4. https://arxiv.org/abs/1803.04585
61. Floridi L, Chiriatti M (2020) GPT-3: Its nature, scope, limits, and consequences. Minds Mach 30(4):681–694. https://doi.org/10.1007/s11023-020-09548-1
62. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI Technical Report. https://cdn.openai.com/better language-models/language_models_are_unsupervised_multitask_learners.pdf
63. OpenAI (2025) HealthBench: A benchmark for evaluating AI systems in health. OpenAI Technical Report. https://openai.com/index/healthbench/
64. Marković, D. & Grollier, J. "Quantum Neuromorphic Computing." Appl. Phys. Lett. 117, 150501 (2020).
65. Wiggers, K. "A Test for AGI Is Closer to Being Solved — but It May Be Flawed." TechCrunch (9 Dec 2024).
66. ARC Prize Foundation. "Announcing ARC-AGI-2 and ARC Prize 2025." https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025 (2025).
67. Bullock, J. B., Hammond, S. & Krier, S. "AGI, Governments, and Free Societies." arXiv:2503.05710 (2025).