The Compound Infeasibility of AGI: Why Replicating and Exceeding Human Cognition Is Not a Scaling Problem*

Solomon Shalom Lijo
Rōmy
solomon@getromy.app

April 28, 2026

Abstract

The dominant formal case against the feasibility of artificial general intelligence (AGI) is complexity-theoretic (van Rooij et al., 2024) and has been answered by the observation that the proof relies on adversarial-distribution assumptions real human behavior does not satisfy (Guerzhoy, 2024). The literature now sits at an impasse: a formal impossibility argument that proves too much, alongside an empirical scaling consensus that 76% of surveyed AI researchers no longer share (AAAI, 2025). We argue this impasse is methodological. The question of AGI feasibility has been treated as single-axis (computational, architectural, or definitional), and single-axis arguments admit single-axis rebuttals. We propose instead a **compound infeasibility thesis** organized around four reinforcing axes—three empirical, one conceptual—none individually fatal, which jointly preclude any plausible path from *current machine-learning paradigms*, on the timeline the inevitability literature describes (single-digit years), to a system matching and exceeding human general intelligence. The operative claim is **infeasibility-on-current-paradigms**, not in-principle impossibility (Section 7.9). The central methodological move is a **cross-axis dependency** (Section 3.5): any plausible mitigation of one axis re-imports demands on at least one other, so single-axis rebuttals do not aggregate into a rebuttal of the conjunction. The four axes are (i) energetic-computational, (ii) developmental and cumulative-cultural, (iii) causal-embodied generalization, and (iv) the absent domain-general specification of *exceeding* human general intelligence. We ground each axis in measured 2024–2026 evidence (BabyLM, the ARC-AGI-2 closure and ARC-AGI-3 launch, GSM-Symbolic, Apple's "Illusion of Thinking," Epoch AI data-exhaustion projections, the AAAI 2025 panel). We engage Hendrycks et al. (2025) directly: the framework substantially advances the matching specification but does not address exceeding, and its own "jagged-profile" diagnosis is the Goodhart pattern our argument predicts. We conclude with three falsifiable refutation conditions: a sample-efficiency threshold, a transfer-without-retraining threshold, and a documented specification-and-realization breakthrough on the exceeding problem. Until any is met, the burden of proof sits with the inevitability claim, not with the skeptic.

Keywords: artificial general intelligence, scaling laws, cognitive science, computational complexity, embodied cognition, cumulative culture, sample efficiency, falsifiability

*Benchmark figures, model scores, and per-task cost estimates cited throughout this paper are time-stamped to April 2026 and may have been superseded by subsequent results. The structural argument is independent of any specific number within plausible ranges; readers should treat the empirical claims in Section 5 as illustrative of the trajectory the compound argument predicts, not as load-bearing on their precise values.

1 Introduction

In March 2026, ARC-AGI-3 launched as a new interactive benchmark for AI reasoning. Within a week, the leading frontier reasoning models—GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro—were scoring near zero on it. Humans (under panel-verified evaluation, multiple humans cross-checking until every task is solved) were scoring 100% (ARC Prize, 2026b; MindStudio, 2026). The benchmark had been launched, in part, because the previous benchmark (ARC-AGI-2) had been closed only weeks earlier by a combination of extended chain-of-thought reasoning, search over candidate solutions, program synthesis, and structured verification (ARC Prize, 2026a; Poetiq, 2026). The closure had required architectural machinery that pure scaling does not produce, and the new benchmark exposed that the closure had not, in fact, generalized to a different format requiring interactive rule discovery. The pattern is not new. It is the dominant data of the field, and it is what the dominant inevitability narrative has been most quietly inattentive to.

Artificial general intelligence (AGI) is the proposition that there exists a buildable artifact whose cognitive performance matches or exceeds human performance across the full range of intellectual tasks. The proposition is defended by frontier laboratories as imminent and, in some treatments, inevitable on a timeline of single-digit years (Shah et al., 2025; Fortune, 2025; Stanford AI Index, 2025). It is contested as impossible by a smaller but increasingly formal literature (van Rooij et al., 2024; Marcus, 2022, 2025; Lake et al., 2017), and as deeply uncertain by the median view of practicing researchers (AAAI, 2025). The disagreement is not merely about timelines. It is about whether the underlying object exists at all, whether it is the kind of thing that admits engineering, and whether the methods currently treated as the path toward it are the methods that would, in principle, get us there.

This paper makes the case that AGI is not feasible on the path the field is currently on, and that the case for infeasibility has been weaker than it should be because the strongest existing argument is structured to invite exactly the rebuttal it has received. We propose a different structure: a compound argument across four independent axes of infeasibility, in which the strength of the case lies not in any single axis but in the conjunction. Each axis is empirically grounded; none requires unfalsifiable distributional assumptions; and, crucially, the argument applies symmetrically to the "replicating" and "exceeding" halves of the AGI proposition, where the existing literature has focused almost exclusively on the former.

1.1 The State of the Debate

The most cited formal argument that AGI is impossible is van Rooij et al. (2024)'s *Reclaiming AI as a Theoretical Tool for Cognitive Science*. The paper formalizes an "AI-by-Learning Problem" and proves that, under their formalization, it is NP-hard under randomized reductions, even granting idealized data and the most efficient machine learning methods. The argument is sometimes summarized as "no amount of compute scaling can get us there," with co-author Olivia Guest framing the resource implication as: "there will never be enough computing power to create AGI using machine learning, because we'd run out of natural resources long before we'd even get close."

The argument has been criticized in detail by Guerzhoy (2024), whose central observation is that the proof equivocates between two distinct interpretations of the data distribution $\mathcal{D}$. The informal target of the argument is the structured distribution of real human situation–behavior pairs (which exhibits hierarchical, causal, and rule-like regularities). But the formal proof works only if $\mathcal{D}$ is permitted to be *arbitrary* (specifically, polytime-sampleable). Once the assumption is made arbitrary, the same proof structure equally implies that learning ImageNet classification is intractable, because ImageNet classification can also be cast in the same complexity-theoretic frame. Since ImageNet is, observably, learnable, at least one component of the proof must be wrong: either the formalization does not capture intractability, or the argument is flawed, or some intractable problem has nevertheless been solved (which is the option no one defends).

The exchange leaves the field in an awkward position. The most rigorous formal argument against AGI's feasibility has a known soft spot. Empirical skepticism is widespread—the AAAI 2025 Presidential Panel survey of 475 AI researchers found that 76% rate it "unlikely" or "very unlikely" that scaling current approaches will deliver AGI (AAAI, 2025)—but empirical skepticism does not, by itself, constitute an argument that the project is infeasible in principle. The result is a literature in which the strongest impossibility argument is rebuttable, the strongest empirical skepticism is non-rigorous, and the strongest inevitability claim faces no formal opposition that survives technical scrutiny.

1.2 Our Contribution

We argue that the impasse is a symptom of single-axis argumentation. van Rooij et al. (2024) attack feasibility from one front (computational complexity); Marcus (2022) attack from a different front (architectural deficiency of pure neural scaling); the embodied-cognition tradition attacks from a third (symbol grounding and sensorimotor experience); the cumulative-culture tradition attacks from a fourth (cognition is not a property of the individual brain). Each of these arguments, as a single-axis case, faces a single-axis rebuttal. The complexity argument has Guerzhoy. The architectural argument has the response that hybrid systems are already incorporating symbolic reasoning. The embodiment argument has the response that LLMs have begun to ground tokens through tool use, code execution, and multimodal training. The cumulative-culture argument has the response that internet-scale corpora are themselves a snapshot of cumulative culture.

The compound argument we propose is that all four constraints obtain simultaneously, that they are not independent (each reinforces the others), and that the partial mitigations against any one of them do not cumulatively bridge the four-axis gap. We make four contributions:

1. We articulate **four axes of infeasibility**—energetic–computational, developmental–cultural, causal–embodied, and specification–target. The first three are grounded in measured empirical quantities rather than adversarial-distribution assumptions; the fourth is conceptual, and we mark the difference in epistemic level explicitly (Section 3).

2. We address the **exceeding problem** as a distinct thesis: even if a system replicated human cognition, the requirement that it *exceed* human general intelligence is, in the existing literature, an unsolved ex-ante engineering specification problem. No domain-general specification of superhuman general intelligence is on offer to optimize against; the available specifications are domain-specific surrogates (benchmark scores, narrow-task superiority) that admit Goodhart-style overfitting and do not aggregate into the target the inevitability claim assumes (Section 4).

3. We **ground the argument in 2024–2026 evidence**: BabyLM sample-efficiency results (Hu et al., 2024), ARC-AGI-2 frontier scores (Chollet et al., 2025), Epoch AI data-exhaustion projections (Villalobos et al., 2024), the AAAI 2025 Presidential Panel report (AAAI, 2025), and brain-energy measurements (Levy and Calvert, 2021). Where existing arguments rely on idealized models, we rely on measurements (Section 5).

4. We propose **three falsifiable refutation conditions** under which the compound argument should be abandoned, and we argue that until any of these conditions is met, the burden of proof sits with the inevitability claim rather than with the skeptic (Section 6).

The operative definition. We adopt, throughout this paper, the dominant industry definition of AGI: *a system that matches and exceeds the cognitive performance of a well-educated adult human across the full range of intellectual tasks, in a way that is domain-general rather than reduced to performance on any particular benchmark or task family.* This is consistent with the framing in Shah et al. (2025), Hendrycks et al. (2025) (whose framework targets the matching

half of this definition specifically), and the broader inevitability literature. Alternative definitions exist (some restrict AGI to "human-comparable performance on a sufficiently broad benchmark suite," which removes the exceeding requirement; others extend AGI to include consciousness or moral status), and we discuss them in Section 8. The compound argument is targeted at the dominant definition. We commit to it here, in the introduction, rather than in a later subsection, so that the argument's scope is explicit before it begins.

What "infeasibility" means in this paper. The word does substantial work in the title and we mean it in a specific scoped sense throughout. The compound argument's operative claim is that AGI under the operative definition is not feasible (a) *on the path described by current machine-learning paradigms*, and (b) *on the timeline the dominant inevitability literature describes* (single-digit years; e.g., Altman (2025); Amodei (2024); Hassabis (2025); Khoja and Hiscott (2025); Shah et al. (2025)). It is not the claim that AGI is in-principle impossible, nor the claim that intelligence-resembling artifacts could never be built by some research program structurally distinct from current paradigms. The two claims dissociate cleanly: a future research program that addresses the four axes through methods qualitatively different from extensions of transformer scaling is not refuted by our argument. We use "infeasibility" rather than "skepticism" or "constraint" because the argument we construct is designed to be load-bearing against the active research program (scale, neurosymbolic patches, agentic tool use, multimodal extension, RL from verifiers), not against a hypothetical future research program of unknown character. The reader who insists on a stronger reading of "infeasibility" is welcome to substitute "compound infeasibility-on-current-paradigms" wherever "compound infeasibility" appears.

What this paper is not. We do not argue that artificial intelligence is a dead end, that LLMs are without value, or that the trajectory of capability progress over the last decade has been illusory. The deployed utility of frontier models in specific contexts (medical question answering (Singhal et al., 2023), legal reasoning (Guha et al., 2024), programming, writing assistance) is real, substantial, and increasing. The argument here is narrower: that the specific proposition labeled "AGI"—under the dominant definition just stated—is not a scaling problem and is not on the path that current paradigms describe. Useful, valuable, even transformative AI systems are entirely compatible with the infeasibility of AGI as defined.

Scope. The contribution of this paper is in the conjunction of four arguments rather than the origination of any single one. We synthesize traditions in cognitive science, computational complexity, embodied cognition, and cumulative-cultural evolution. Each tradition has its own primary literature, which we cite; the structural claim is that the arguments are stronger together than separately, and that the partial mitigations against any one of them do not aggregate. We therefore describe the paper as a synthesis-with-novel-conjunction rather than as an originating contribution to any of the four constituent traditions.

1.3 Roadmap

Section 2 surveys the existing impossibility arguments and their respective rebuttals. Section 3 presents the four axes of compound infeasibility and (in Section 3.5) develops the cross-axis dependency that does the structural work of the compound thesis: four propositions establishing that any plausible mitigation of one axis re-imports demands on at least one other. Section 4 develops the exceeding problem in detail. Section 5 grounds the argument in 2024–2026 empirical data, including the ARC-AGI-2 closure and ARC-AGI-3 launch. Section 6 states the refutation conditions. Section 7 engages the strongest counterarguments. Section 8 states the argument's limitations. Section 9 concludes.

2 Existing Impossibility Arguments and Their Limits

We survey four families of existing arguments against AGI feasibility and identify, in each case, the rebuttal that has limited the argument's force. The goal is not to dismiss these arguments but to locate the structural feature—single-axis presentation—that exposes them to those rebuttals, so that the compound argument we develop in Section 3 can be designed to avoid the same vulnerability.

2.1 The Complexity-Theoretic Argument

van Rooij et al. (2024) present the most rigorous formal case. Their AI-by-Learning Problem asks: given a sample from a distribution $\mathcal{D}$ over situation–behavior tuples (s_i, b_i), find an algorithm of bounded description length that maps situations to behaviors with accuracy exceeding chance by at least $\varepsilon(n)$, with probability at least $1 - \delta(n)$. They prove that, under their formalization, the problem is NP-hard under randomized reductions, even given idealized resources. The intended interpretation is that human-level cognition cannot be obtained by machine learning, full stop.

The Guerzhoy (2024) critique identifies that the proof's reduction works because $\mathcal{D}$ is permitted to be arbitrary (any polytime-sampleable distribution), whereas the informal claim depends on $\mathcal{D}$ being the *actual* distribution of human situation–behavior pairs. The actual distribution is structured: it satisfies hierarchical regularities, causal constraints, language statistics, embodiment constraints, and an enormous number of empirical regularities that distinguish human behavior from arbitrary input–output tuples. The proof works only because the formalization throws away that structure.

The same proof structure, applied to ImageNet, would reach the (false) conclusion that ImageNet classification is intractable, because ImageNet's natural-image distribution can also be embedded in the same arbitrary-distribution frame. Since ImageNet is observably tractable, the proof must be flawed at the formalization step. Specifically, the proof ignores that real-world learning algorithms have inductive biases (convolutional architectures match natural-image statistics; transformers match sequence dependency structure) that make structurally complex distributions efficiently learnable when their structure is matched.

The lesson for our argument is that any formal infeasibility claim that depends on adversarial-distribution assumptions will be vulnerable to the same critique. We design our argument around *measured* properties of human cognition (sample efficiency, transfer, energetic cost, cultural transmission) rather than around idealized worst-case distributions.

2.2 The Architectural-Deficiency Argument

Marcus (2022) argued in 2022 that pure scaling of large language models would not deliver AGI, that "scaling laws" were empirical generalizations rather than physical laws, and that the path forward required hybrid neurosymbolic architectures combining the pattern-recognition strengths of deep learning with the structured reasoning of symbolic systems. The argument has aged well: Marcus (2025) reviews the 2024–2025 emergence of neurosymbolic methods (DeepSeek R1's rule-based reward verification, OpenAI o3's structured reasoning chains, the explicit incorporation of program-synthesis and verification components in frontier reasoning systems) as substantial vindication of the original critique.

The architectural argument's limitation is that it identifies a deficiency of one paradigm rather than an in-principle barrier. The neurosymbolic response is precisely that the deficiency is architectural and therefore architecturally addressable: if the missing piece is symbolic reasoning, add symbolic reasoning; if the missing piece is causal modeling, add causal modeling. As long as each individual deficiency admits a fix, the architectural argument supports a critique of *current* systems, not an infeasibility claim about AGI in general.

The compound argument we develop accepts the architectural-addressability response on its own terms but argues that it does not aggregate to AGI. The reason—developed formally in Section 3.5 as the cross-axis dependency—is that each architectural fix that closes one axis hardens or re-imports demands on at least one other axis. The empirical signature is visible in the very ARC-AGI-2 closure that vindicates Marcus (Sections 3.3 and 5.2): the closure was real, the architectural machinery (extended chain-of-thought, search over candidates, program-synthesis components, verification) was the mechanism, and pure scaling did not produce it. But the closure consumed two-to-three orders of magnitude more per-task compute than humans use on the same problems (worsening Axis I), it was format-specific and did not transfer to ARC-AGI-3 (Axis III remained open under a new format), and the closure did not constitute progress toward a domain-general specification of exceeding (Axis IV remained untouched). *Architecturally addressable* is therefore not the same as *aggregately addressable*, and the compound argument is targeted at the latter. The reader who notices that Section 3.3 endorses Marcus's empirical vindication while the present subsection limits Marcus's architectural argument is reading both passages correctly: the reconciliation is in the cross-axis dependency, not in either standalone passage.

The lesson for our argument is that an axis of infeasibility under the compound thesis must be one for which the architectural fixes do not in fact close the gap, either because the gap is multidimensional (closing one dimension does not address the others) or because the fix imports a new constraint of equal severity. The energetic, developmental, and grounding axes we develop below have this property in a way that "add symbolic reasoning" does not.

2.3 The Symbol-Grounding and Embodiment Arguments

The symbol grounding problem (Harnad, 1990, 1993) holds that a purely symbolic system cannot have semantics in any non-derivative sense, because its symbols' connections to the world are mediated only by external interpreters. The frame problem (Dennett, 1984) is the practical face of the same issue: a symbol system without grounding cannot anticipate which contingencies are relevant in any given situation. The embodied cognition tradition (Lakoff and Johnson, 1999; Clark, 2008; Chemero, 2009) extends the argument: cognition is not merely brain-bound, but constituted by an agent's sensorimotor engagement with its environment.

Recent treatments (Pavlick, 2023; Farkaš et al., 2025) have complicated the picture for LLMs specifically. Token-level operations on internet-scale corpora are not, narrowly, symbol manipulation in the classical sense; the corpora themselves contain enormous embodied-experiential residue (descriptions of sensory experience, action sequences, causal narratives), and multimodal training extends the grounding surface. Harnad (2024) argues that the partial successes of LLMs reveal that symbol grounding is more graded than the original formulation supposed, with formal and mathematical structures occupying the least-grounded end of the spectrum.

The lesson for our argument is that "LLMs lack grounding" is not a clean axis of infeasibility on its own; the more defensible claim is that LLMs lack the specific kind of *causal* grounding that supports human generalization to genuinely novel situations, and that scaling token counts on text corpora does not produce that kind of grounding. We sharpen this into the causal-embodied axis in Section 3.3.

2.4 The Cumulative-Culture Argument

Tomasello (1999) and Henrich (2015) develop the case that human cognition is not primarily a property of the individual brain but a cumulative-cultural artifact: humans are not, in many measures of fluid intelligence, cognitively superior to other great apes (Henrich, 2015; Heyes, 2018), and the human capacity for abstract reasoning is bootstrapped over evolutionary and cultural timescales by social learning, language, joint attention, and what Tomasello (1999)

calls the ratchet effect—the high-fidelity transmission and incremental modification of cultural innovations across generations.

The argument's implication for AGI is that a system trained on a static corpus is a system trained on a snapshot of cumulative culture, not a system embedded in the cumulative-cultural process. The cumulative-culture critique of LLMs is sometimes dismissed on the grounds that internet-scale corpora are precisely a record of cumulative culture, which is a partial answer: the corpora are a snapshot, not a process. The corpora cannot, by their nature, contain the future cultural innovations that will be produced by humans embedded in the process; they contain only the innovations that have already occurred and have been recorded.

The lesson for our argument is that cumulative culture is not a feature that can be added to an LLM by scaling, because the relevant property is dynamic embedding in the social-learning process, not static representation of its outputs. We develop this into the developmental–cultural axis in Section 3.2.

2.5 Why the Existing Arguments Have Not Settled the Question

Each of the four arguments above identifies a real problem. None has settled the question of AGI feasibility because each, taken alone, admits a partial mitigation that the inevitability literature treats as a complete rebuttal. The complexity argument admits the inductive-bias mitigation. The architectural argument admits the neurosymbolic mitigation. The grounding argument admits the multimodal mitigation. The cumulative-culture argument admits the snapshot-corpus mitigation. The compound argument we develop next is that the partial mitigations against any one of these arguments do not cumulatively address all four, and that the constraints reinforce rather than substitute for one another.

3 The Compound Infeasibility Thesis

We propose four reinforcing axes of infeasibility. The four are not mutually exclusive with the existing literature; each draws on prior work in the corresponding tradition. The novelty is in their conjunction, in the choice to ground each axis in measured rather than idealized quantities, and in the explicit treatment of how each constraint reinforces the others. We state the thesis formally and then develop each axis in turn.

Claim 1 (Compound Infeasibility)**.** *A system that matches and exceeds human general intelligence across domains is not feasible on the path described by current machine-learning paradigms, because:*

(i) the energetic and computational economy of human cognition is several orders of magnitude beyond the silicon scaling curves observable through 2026, and the only known directions to closing that gap require substrate changes that re-import the remaining axes' constraints (Section 3.1);

(ii) human cognition runs on developmental and cumulative-cultural scaffolding that contemporary ML pipelines do not replicate (Section 3.2);

(iii) human generalization across genuinely novel domains depends on causal-embodied grounding that scaling on static corpora does not produce (Section 3.3);

(iv) the existing literature has produced no domain-general specification of exceeding *human general intelligence; the available specifications are domain-specific surrogates that admit Goodhart-style saturation, and the production of a domain-general specification is a nontrivial conceptual open problem rather than an engineering side-effect of closing axes (i)–(iii) (Section 3.4, developed in Section 4).*

The four constraints are not independent. Section 3.5 establishes a cross-axis dependency*: any plausible mitigation of one axis re-imports demands on at least one other axis, so the partial rebuttals to each single-axis argument do not aggregate to bridge the joint gap.*

3.1 Axis I: The Energetic and Computational Economy of the Brain

The human brain operates on approximately 20 watts (Levy and Calvert, 2021). The cerebral cortex specifically consumes roughly 5 W of that total budget; within the cortex, only about 0.1 W of ATP-equivalent power is partitioned to computation, while ∼3.5 W—35-fold more than computation—is consumed by communication via axonal action potentials and synaptic transmission (Levy and Calvert, 2021). The remainder of the 20 W brain budget is consumed by subcortical structures, white-matter signalling, and metabolic maintenance (Karbowski, 2009). The brain executes on the order of 10^{14}–10^{16} synaptic operations per second on this budget, with average energy per synaptic event in the 10^{-14} to 10^{-15} J range (Levy and Calvert, 2021). There are $\sim 86 \times 10^9$ neurons (Azevedo et al., 2009) and $\sim 1.5 \times 10^{14}$ synapses (Drachman, 2005), organized in a topology that does not correspond cleanly to the dense matrix multiplications that dominate transformer compute.

For comparison: training GPT-4-class models has been estimated to consume on the order of 5×10^4 MWh of energy (Patterson et al., 2022; Epoch AI, 2025), with subsequent generations consuming substantially more. Inference at scale (GPT-5 serving ∼2.5 billion prompts per day) consumes on the order of 850 MWh per day, dominated increasingly by reasoning-chain compute rather than by the original forward pass (Epoch AI, 2025). The 2025 ratio of frontier-model training-and-inference energy to brain energy budget is on the order of 10^8 for inference alone—and this ratio understates the gap because it compares one human's brain to the joint energy cost of serving all users. Per-task energy comparisons are even more unfavorable: the brain solves novel reasoning tasks (e.g., the kind of abstraction problems posed by ARC-AGI-2 (Chollet et al., 2025)) at human-typical ∼60–80% accuracy on a per-task energy budget measured in joules. Frontier reasoning models in 2025–2026 score 9.9% (GPT-5) to 37.6% (Opus 4.5 with 64k thinking tokens) on the same benchmark, at per-task costs of $0.73–$2.20 corresponding to substantially more joules of compute (Chollet et al., 2025; ARC Prize, 2025).

The standard response to the brain's energy efficiency is that architectural improvements will close the gap: neuromorphic computing, specialized hardware, and algorithmic efficiency gains are all observably trending in the right direction. We grant the trend. The argument is not that the gap is unclosable in principle but that the magnitude of the gap, combined with the rate at which the silicon-side trend is decelerating, makes "close it via scaling" a quantitatively implausible plan. Aggregate global data-center electricity consumption was 5.2% of U.S. total in 2025 and is projected to reach 8–10% by 2030 (IEA, 2025); the rate of efficiency improvement on silicon is roughly 1.3× per generation (Thompson et al., 2020); the energetic gap to be closed is on the order of 10^6–10^8 depending on how it is measured. The arithmetic does not work even on the most optimistic decade-scale projections.

Existence vs. deployment. A standard objection is that AGI does not need to be *energy-efficient* to exist—a system that performs the relevant cognitive work at $10^6\times$ the brain's energy budget is still AGI under the operative definition. We accept that distinction sharply. The axis is not the claim that AGI must run at 20 W or it is not AGI; it is the claim that AGI under the operative timeline (single-digit years, on the path of current paradigms) cannot run at the scale the inevitability claim describes without resolving the energetic constraint, and the only known directions to closing the constraint are substrate changes that re-impose the other three axes. Three observations make this concrete. First, the inevitability claim is not about an isolated artifact in a research lab; it is about a deployable system, with serving costs that scale with usage (Epoch AI, 2025). The deployment-economics constraint is therefore not an aesthetic preference for efficiency—it is part of what "feasible on the path of current paradigms" means, because

current paradigms are productized via at-scale inference and a 10^8 ratio against human energy costs is incompatible with the deployment surface the inevitability literature assumes. Second, the directions known to close the energetic gap by orders of magnitude—spiking neural networks on neuromorphic hardware (Loihi, BrainScaleS, SpiNNaker), in-memory computing, photonic processors—are substrate changes, not algorithmic refinements within transformers; they require retraining paradigms that current corpora and inductive biases do not match, and they re-import the developmental and grounding problems we develop in Sections 3.2 and 3.3. Third, even granting the strong existence reading (a working AGI at $10^6\times$ brain energy), the existence reading dissociates from the operative claim: the operative claim is that AGI is not feasible *on the path of current paradigms* on a single-digit-year timeline, and the energetic constraint cuts against feasibility-on-current-paradigms even if it does not cut against in-principle existence.

The energetic axis is therefore not, on its own, an in-principle infeasibility argument. The axis becomes load-bearing in the compound argument because (i) under current paradigms, the gap directly constrains deployment-scale feasibility on the operative timeline, and (ii) the kind of breakthrough that would close the gap (e.g., a switch to spiking neural networks running on truly neuromorphic hardware) is also the kind of breakthrough that would force the system to confront the developmental, embodiment, and cumulative-culture problems we develop next, none of which is addressed by hardware change. The cross-axis dependency is rendered formally in Section 3.5.

3.2 Axis II: Developmental and Cumulative-Cultural Scaffolding

A 13-year-old human is exposed to at most $\sim 10^8$ words of language input, an upper bound derived by the BabyLM Challenge (Hu et al., 2024) from longitudinal studies of child language exposure (Hart and Risley, 1995; Roy et al., 2015). Frontier LLMs are trained on 10^{12}–10^{13} tokens—four to five orders of magnitude more linguistic input than any human will ever encounter (Hoffmann et al., 2022; Villalobos et al., 2024) (with the caveat, developed below and revisited in Section 5.1, that the cross-substrate comparison is not pure tokens-to-tokens: human linguistic exposure is embedded in multimodal embodied experience and cumulative-cultural scaffolding that LLM training is not). The disparity is the central motivating finding of the BabyLM line of research: at the data scale matched to a 13-year-old's lifetime input, the best 2023 and 2024 submissions to the 100M-word track (Hu et al., 2024) exhibit a substantial competence gap relative to the same architectures trained at trillion-token scale on standard downstream evaluations. The gap has narrowed over the two challenges—some 2024 entries close meaningfully on specific subtests—but remains large in aggregate, and the children whose input distribution the dataset is matched to do master their native languages on that budget. The architectural gap, not children themselves, is the relevant comparison point; the LM-vs-LM gap measured by the challenge stands in for the broader claim that humans are sample-efficient at a level no current architecture approaches.

The disparity is not addressed by saying that humans inherit evolutionary priors. The standard response that "humans have ancestral genetic information from 10,000+ generations" (Mu, 2025) acknowledges the existence of inductive bias without quantifying it, and the relevant inductive biases are not architectural in the transformer-architecture sense; they are developmental. A child is embedded in a social environment with joint attention (Tomasello, 1999), ostensive communication (Csibra and Gergely, 2009), scaffolded teaching (Vygotsky, 1978), and feedback loops between action, prediction, and observation that no LLM training pipeline replicates. Children acquire causal models of the physical world before they acquire much language at all (Spelke and Kinzler, 2007); they acquire theory-of-mind through embodied participation in social interaction (Frith and Frith, 2008); they acquire counterfactual reasoning capacities that LLMs systematically fail to exhibit even at frontier scale (Pearl and Mackenzie, 2018; Bender et al., 2021).

The cumulative-cultural dimension extends the disparity. Even the cognitive capacities that look most like "raw intelligence"—abstract symbolic reasoning, mathematics, formal logic—are

not properties of the individual brain in the way that vision or motor control are. They are cognitive gadgets, in the sense of Heyes (2018): capacities assembled from cultural inheritance, transmitted with high fidelity by social learning and explicit instruction, ratcheted across generations (Tomasello, 1999). The implication is that the substrate on which human general intelligence runs is not the brain alone; it is the brain embedded in a multi-generational social-learning process. The relevant comparison is not "one human brain versus one trained model" but "the cumulative inheritance of ∼200,000 years of cumulative culture, transmitted through high-fidelity teaching and language, scaffolded by embodied social interaction, versus an LLM trained on a static corpus."

The static-corpus mitigation does not work, because the relevant property is dynamic embedding rather than recorded output. Henrich (2015) makes the point sharply: an individual human, raised without cultural input, is a cognitively unimpressive primate. The capacity for general intelligence is not in the head; it is in the head's dynamic relationship to a multi-generational social-learning process. A system that learns from a snapshot of that process can recapitulate, at best, the patterns visible in the snapshot; it cannot extend the process, because it is not embedded in the process.

3.3 Axis III: Causal-Embodied Grounding and Generalization

The third axis is generalization to genuinely novel domains. Human cognition does this routinely: a child who has learned to navigate physical objects can reason about novel objects she has not seen (Spelke and Kinzler, 2007); an adult who has learned one programming language can transfer skills to a structurally different one; a researcher in one field can recognize when a result from another field is relevant. The capacity rests, on the embodied-cognition account, on *causal* models of the world that are constructed through sensorimotor interaction: the agent learns what happens when she pushes, pulls, drops, combines, or transforms objects, and these interactions seed the abstract relational schemas that later support transfer (Lakoff and Johnson, 1999; Clark, 2008).

LLMs, trained on text, learn statistical regularities of co-occurrence rather than causal regularities. The distinction is well-established (Pearl and Mackenzie, 2018): correlation is not causation, and statistical models trained on observational data without intervention do not, in general, recover the causal structure that generates the data. Multimodal training (image, video, audio) extends the surface of the data but does not, by itself, introduce intervention; an LLM that has seen a billion videos of objects falling has not dropped a single object. The recent line of work on agentic models—LLMs that take actions in the world via tool use, code execution, and environment interaction—begins to introduce intervention, and is the most plausible direction for closing this gap (Brohan et al., 2023; Wang et al., 2024). But the actions an agentic LLM takes are bounded by the tool surface available to it, and the breadth of that surface remains far short of the open-ended sensorimotor engagement available to a human child.

The empirical consequence is visible on benchmarks specifically designed to measure causal-grounded generalization, and the trajectory of these benchmarks is itself revealing. The ARC-AGI program is grounded in Chollet (2019)'s definition of intelligence as skill-acquisition efficiency on novel tasks under bounded experience, rather than as performance on any pre-specified task domain—a definition that, for our purposes, is the cleanest extant formalization of the causal-grounded generalization capacity Axes II and III target. ARC-AGI-2 (Chollet et al., 2025) operationalizes this definition: the solver must infer abstract transformation rules from a small number of input–output examples and apply them to held-out test cases. Through 2024 and into early 2025, frontier-model performance on ARC-AGI-2 remained well below human levels: GPT-5 scored 9.9%, Opus 4.5 reached 37.6%, and bespoke refinement systems built on Gemini 3 Pro reached 54% at $30/task (ARC Prize, 2025). By early 2026, however, both Gemini 3 Deep Think and GPT-5.5 have crossed 84–85% on ARC-AGI-2 (ARC Prize, 2026a; Poetiq, 2026), with per-task costs of roughly $1.87–$13.62. ARC-AGI-2 has, in this sense, been closed.

The closure does not undermine the compound argument. It illustrates one of its central mechanisms. The architectures that closed ARC-AGI-2 are explicitly hybrid: extended chain-of-thought reasoning, search over candidate solutions, program synthesis components, and structured verification of intermediate steps. The closure required precisely the architectural shift that Marcus (2025) predicted scaling alone would not produce. And the closure has been answered, almost immediately, by the launch of ARC-AGI-3 (ARC Prize, 2026b) on March 25, 2026, an interactive benchmark in which agents must discover rules and goals in video-game-like environments without explicit instructions. As of April 2026, GPT-5.4 scores near zero, Claude Opus 4.6 scores 0.25%, and Gemini 3.1 Pro scores 0.37% in their default deployed configurations; humans, under panel-verified evaluation in which multiple humans cross-check until every task is solved, reach 100% (MindStudio, 2026). The closure mechanisms that have made the most progress are bespoke architectural harnesses on top of frontier LLMs and purpose-built non-LLM agents—both of which outperform default-configuration frontier LLMs by one to two orders of magnitude, and both of which import architectural machinery the default configurations lack. Symbolica's Arcgentica (Symbolica, 2026), an orchestrator–subagent program-synthesis harness *built on Claude Opus 4.6*, achieves 36.08% (113 of 182 playable levels, 7 of 25 games) at a full-run cost of approximately $1,005, compared with Opus 4.6 in its default configuration scoring 0.25% at approximately $8,900 for an equivalent run. This is the central observation, and it sharpens the case rather than weakens it: *the same underlying model* goes from 0.25% to 36.08% at one-tenth the cost when given the right architectural harness. The capability gain is not in the model's parameters or training data; it is in the orchestration layer on top of them. This is direct empirical support for the architectural-addressability response (Section 2.2) and the cross-axis dependency (Section 3.5): scaling alone is not what closes Axis III, bespoke architectural machinery is, and the machinery is format-specific to ARC-AGI-3. The non-LLM entries StochasticGoose (Tufa Labs, a CNN with reinforcement learning predicting frame changes, 12.58%) and Blind Squirrel (an explore-and-learn state-graph agent, 6.71%) likewise outperform default-configuration frontier LLMs by an order of magnitude, demonstrating that even non-LLM primitives are better matched to the format than chat-configured transformers. The 100-percentage-point gap between panel-verified humans and default-configuration frontier LLMs is the current state of the causal-grounded generalization gap.

The benchmark-closure pattern is itself the data. Each generation of benchmarks specifically designed to test causal-grounded generalization is closed, eventually, by the addition of architectural machinery that current paradigms lack at training time and import at inference time. Each closure prompts the design of a new benchmark that exposes the next layer of the gap. The pattern is consistent with the compound argument: the architectural fixes do not address the underlying constraint (sample-efficient causal learning grounded in embodied interaction), they substitute increasingly expensive search-and-verification for the missing capability, and the next benchmark exposes that the substitution has not, in fact, generalized.

The two readings of the closure pattern. The pattern admits two interpretations and we owe the reader an argument for the one we adopt. Reading A, which we defend: the underlying capability of causal-grounded generalization does not advance proportionately with benchmark scores; each closure imports format-specific architectural machinery whose effect is to substitute search and verification for the missing capability, and the format-specific machinery does not transfer when the format changes. Reading B, the inevitability-friendly alternative: each new benchmark is genuinely harder than the last, the closures reflect real capability advances, and the appearance of recurring gaps is an artifact of the field's habit of designing harder probes whenever the previous probe falls. Reading B is the move that skeptical literature has historically had to defend against (when ImageNet fell, when Go fell, when GLUE fell); we owe a discriminator. The discriminator, on the available evidence, is whether the architectural machinery that closes benchmark N transfers, with comparable efficacy, to benchmark $N{+}1$. Under Reading B, partial

transfer is expected: the same capability that closed N should help on $N{+}1$, even if $N{+}1$ is harder. Under Reading A, transfer is minimal: the closure mechanism is benchmark-specific machinery, not capability progress, so a different benchmark format requires different machinery built largely from scratch. The empirical fact is that on the day ARC-AGI-3 launched (March 25, 2026), the same frontier reasoning models that had recently crossed 84–85% on ARC-AGI-2 scored 0–0.37% on ARC-AGI-3 in their default configurations (MindStudio, 2026; ARC Prize, 2026a); the only systems making substantial progress were bespoke architectural harnesses (Symbolica's Arcgentica—an Opus-4.6-on-program-synthesis-harness—at 36.08%) and purpose-built non-LLM agents (Tufa Labs' CNN-plus-RL StochasticGoose at 12.58%, the state-graph Blind Squirrel at 6.71% (Symbolica, 2026)), each outperforming default-configuration frontier LLMs by one to two orders of magnitude. The architectural primitives that closed ARC-AGI-2 are not even directionally well-suited to ARC-AGI-3 in the deployed transformer configuration; closing the gap on ARC-AGI-3 has required either swapping in an entirely different architectural primitive (StochasticGoose, Blind Squirrel) or wrapping the existing LLM in a benchmark-specific orchestrator (Arcgentica). Either way, the transfer that Reading B predicts is absent at the magnitudes that would matter. We grant Reading B is partially correct—each benchmark does measure something new by construction—but the empirical signature on transfer is more consistent with Reading A's structural claim than with the strong form of Reading B. The compound argument is robust to a mixed reading: even granting that some capability progress is real (Reading B partially correct), the format-specific cost of each closure and the absence of cross-benchmark transfer is itself the cross-axis dependency on Axes I and III—the kind of evidence the structural argument predicts and the inevitability claim does not.

The relevant generalization measure is therefore not score-on-current-benchmark but the time and compute required to close each successive benchmark, the architectural complexity required, and whether the closure transfers to the next benchmark in the sequence. The recent empirical literature converges on a consistent picture. Lewis and Mitchell (2024) demonstrate that human performance on analogical reasoning is robust to counterfactual variants of the same problems, while LLM performance drops sharply when the surface form is altered, even when the underlying abstract structure is preserved. Mirzadeh et al. (2024), in the GSM-Symbolic benchmark from Apple, show that LLM accuracy on grade-school mathematics deteriorates as problem complexity grows and that, on the GSM-NoOp variant (which adds a seemingly relevant but operationally irrelevant clause to each problem), all evaluated models drop substantially—with smaller open-weight models such as Phi-3-mini falling by up to $\sim$65%, and frontier proprietary models still dropping by tens of percentage points. The authors conclude that "current LLMs cannot perform genuine logical reasoning; they replicate reasoning steps from their training data." Kambhampati et al. (2024) provides systematic evidence that the apparent reasoning abilities of LLMs are dominated by exemplar-query similarity and approximate retrieval rather than systematic reasoning, and that LLMs cannot, on their own, do planning or self-verification. Apple ML Research (2025) extends the analysis to large reasoning models with extended chain-of-thought and finds that reasoning effort scales with complexity up to a threshold, then declines despite available token budget, with both standard and reasoning-augmented models collapsing on high-complexity problems. The conclusion across these four lines of work is consistent with Lake et al. (2017)'s earlier framing: building machines that learn and think like people requires causal models, intuitive theories of physics and psychology, and compositional generalization—capacities that scaling on observational data has, on the available benchmark trajectory, not produced.

The causal-grounded gap is not closed by adding more training data, because the limitation is not in the breadth of the corpus but in the kind of regularities the model is exposed to. It is not closed by adding more compute, because compute scaling on text corpora does not introduce intervention. It is closed, in principle, by genuinely embodied agentic systems that learn through interaction—which returns us to the energetic axis (intervention is expensive, in robots even more than in simulation), the developmental axis (the timeline for an embodied agent to acquire

human-comparable causal models is on the order of human childhood years, not training run hours), and—the next axis—the question of what the system is supposed to do once it matches human performance.

3.4 Axis IV: The Specification–Target Problem (Preview)

The fourth axis is structurally distinct from the first three. Axes I–III are anchored in measured empirical quantities (energetic budgets, token counts, benchmark scores), and the gaps they identify can be widened or narrowed by future measurement. Axis IV is conceptual: the requirement that AGI *exceed* human general intelligence faces, on inspection, an open specification problem rather than a measurable shortfall. We treat it as a co-pillar of the compound argument because the conceptual axis interlocks with the empirical axes (closing the energetic, developmental, and causal-grounding gaps does not, by itself, supply a coherent target for "exceeding"), but we mark the difference in epistemic level here so that the reader can hold the two kinds of constraint separately.

A clarification on what the axis is and is not. The axis is not the claim that humans cannot, in principle, build something that exceeds them in some domain—the existence of chess engines, image classifiers, and protein-folding predictors falsifies the strong reading immediately. Nor is it the claim that exceeding could never be *observed*; an ex-post observation that a system has, in some loose sense, exceeded humans is consistent with the axis. The axis is specifically that *ex-ante engineering specification of a domain-general exceeding target*—the kind of specification an inevitability claim that frames AGI as engineerable in single-digit years would require—has not been produced and faces structural obstacles that we develop in Section 4. The distinction matters because the inevitability literature treats AGI as an engineering target on a near-term timeline; engineering targets need ex-ante specifications, not just ostensive ones, and this is where the existing literature is silent about exceeding.

The exceeding requirement: existing impossibility arguments focus almost exclusively on the replication target (can a system be built that matches human cognition?). The exceeding requirement faces a separate problem: even granting replication, the domain-general specifications that would license a claim of "exceeding general intelligence" are absent from the literature. Specifications that are workable are domain-specific (chess Elo, ARC accuracy, mathematical olympiad performance), and domain-specific surrogates admit Goodhart-style overfitting in which the surrogate is saturated without the underlying capability appearing. We develop the conceptual argument at length in Section 4; we flag it here as the fourth pillar of the compound argument and as the axis where existing impossibility literature is most underdeveloped.

3.5 Cross-Axis Dependency: Why the Mitigations Do Not Aggregate

The compound argument's central methodological move is the cross-axis dependency: the partial mitigations against each single-axis argument do not aggregate to a rebuttal of the conjunction. The strength of the compound thesis rests on this structural claim, and we make it explicit here—the previous gestural treatment ("each amplifies the others") is the place reviewers should and do press, and we owe a more rigorous account.

The argument is structural, not a formal proof. We articulate it as a definition and four propositions; each proposition is in principle contestable, and we attach a specific refutation condition to each. The compound thesis is load-bearing if the propositions hold; it is weakened in proportion to how many fail.

Definition 1 (Cross-Axis Dependency)**.** *A set of constraints $\{C_1, \ldots, C_n\}$ exhibits cross-axis dependency if, for every C_i and every plausible mitigation M_i of C_i, there exists some C_j $(j \neq i)$ such that the application of M_i either (a) increases the demands C_j imposes on the candidate solution, or (b) requires that C_j be addressed as a precondition.*

We claim that the four axes of Claim 1 exhibit cross-axis dependency. Without dependency, the four axes would be independently mitigable and the compound argument would reduce to "compound skepticism"—a conjunction of single-axis difficulties, each separately addressable. With dependency, the conjunction has structural force the disjunction of single-axis arguments lacks. We articulate the dependency as four propositions, one per axis, each stating a specific way that closing that axis re-imports demands on another.

Proposition 1 (Energetic-Axis Dependency)**.** *Any plausible mitigation of Axis I (closing the energetic gap by orders of magnitude on the operative timeline) requires substrate or paradigm changes that re-impose demands on Axis II or III.*

The known directions for orders-of-magnitude reduction in per-task compute are substrate changes: spiking neural networks on neuromorphic hardware (Loihi, BrainScaleS, SpiNNaker), photonic processors, in-memory computing, biological substrates. None is a refinement of transformer training that preserves the existing data, training paradigm, or evaluation pipeline. Each requires retraining from primitives that current text corpora do not match (Axis II): spiking networks are not trained by next-token prediction, neuromorphic hardware does not run dense matrix multiplication efficiently, photonic processors do not implement the same operator stack. Substrate-preserving algorithmic improvements are observed on the order of 1.3× per generation (Thompson et al., 2020), far below the rate required to close the 10^6–10^8 gap on a single-digit-year timeline. The proposition is refuted if a substrate-preserving algorithmic improvement of $\sim 10^4\times$ in per-task compute is demonstrated on a representative reasoning benchmark over the next decade.

Proposition 2 (Developmental-Axis Dependency)**.** *Any plausible mitigation of Axis II (closing the sample-efficiency and cumulative-cultural gap by orders of magnitude) requires inductive biases that themselves require Axis III's closure as a precondition.*

The kind of inductive biases that would let a system learn at adult-human linguistic competence from 10^8 words or fewer are biases that match the structure of a mind embedded in social-physical interaction: causal models of physical objects, intuitive physics, theory of mind, joint attention, ostensive pragmatics, scaffolded teaching. These biases are not architectural in the transformer-architecture sense; they are constituted by embodied developmental experience and cumulative-cultural inheritance. The BabyLM line of work (Hu et al., 2024) reports that two years of community optimization on the 100M-word constraint has not produced an architecture matching the same architecture's competence at trillion-token scale; the gap has narrowed but remains large. The absence of the relevant inductive biases is the most parsimonious explanation. Closing Axis II therefore requires Axis III, not as a side effect but as a precondition: the inductive biases that would license sample-efficient learning are precisely the causal-embodied biases that text-only training does not produce. The proposition is refuted if a transformer-architecture submission to the BabyLM 100M-word challenge demonstrates adult-human-comparable linguistic competence on a defensible aggregate of measures (BLiMP, GLUE, EWoK), without auxiliary embodied or multimodal supervision.

Proposition 3 (Causal-Embodied-Axis Dependency)**.** *Any plausible mitigation of Axis III (closing the causal-embodied generalization gap on novel-domain transfer) under current architectural primitives re-imports Axis I; if mitigated through simulation only, it re-imports the synthetic-data ceiling that Axis II warns against.*

The post-2024 embodied wave—OpenVLA (Kim et al., 2024), π_0 (Black et al., 2024), RT-2 (Brohan et al., 2023), Genie 2 (Parker-Holder et al., 2024)—introduces intervention surfaces that text-only training lacks. We treat the mitigation seriously (Section 7.8). The proposition's force is in the dependency: real-world intervention (robotics) is energetically more expensive than text scraping by orders of magnitude per useful sample, re-imposing Axis I; simulation-based

intervention (Genie 2-style synthetic worlds) does not introduce true causal regularities, only those the simulator captures, so it re-imposes the synthetic-data ceiling (Shumailov et al., 2024; Gerstgrasser et al., 2024) which is itself an instance of Axis II's underlying constraint. Closure of Axis III at the scale and breadth required for human-comparable causal-grounded generalization therefore implies either (a) a real-world intervention regime whose energetic cost is incompatible with current-paradigm deployment, or (b) a simulation regime that bottoms out on the simulator's fidelity. The proposition is refuted if an embodied agentic system, trained at energy and data costs comparable to current LLM training, achieves human-comparable causal-grounded generalization on a contamination-resistant out-of-distribution benchmark (e.g., an ARC-AGI-3 successor).

Proposition 4 (Specification-Axis Independence)**.** *Mitigations of Axes I, II, and III do not, even cumulatively, address Axis IV. The production of a domain-general specification of* exceeding *human general intelligence is conceptually orthogonal to engineering closure of the empirical axes, and the existing matching specifications (e.g., Hendrycks et al. (2025)) are scoped by construction to matching, not to exceeding.*

Suppose Axes I–III were closed in some plausible engineering trajectory: the system that emerges is a sample-efficient, energetically-tractable, causally-grounded artifact whose performance can be evaluated against any specific human-comparable benchmark. The exceeding question—what does it mean to optimize for "general intelligence above human level"—remains open. Existing matching specifications are constructed by adapting human-calibrated psychometric frameworks (Cattell–Horn–Carroll); the calibration is intrinsically bounded by the human population the framework was designed against. "200% on the Hendrycks score" is not a defined object on the framework's own terms. Domain-specific surrogates (chess Elo, ImageNet accuracy, ARC score) admit Goodhart-style saturation under sustained optimization (Goodhart, 1975; Strathern, 1997; Gao et al., 2022), and an aggregate that escapes Goodhart in the exceeding regime is a problem the existing literature has not solved (Section 4). The proposition is refuted by the publication of a domain-general specification of exceeding satisfying the criteria of Section 4.4, accepted by a representative cross-section of the AI research community, and empirically realized.

Synthesis. The four propositions describe a coupled mitigation graph. Axis I's mitigations re-impose Axis II (and possibly III). Axis II's mitigations require Axis III as a precondition. Axis III's mitigations re-impose Axis I (or, on simulation, Axis II). Axes I, II, III's mitigations do not, even jointly, address Axis IV. The empirical axes form a cycle (I $\leftrightarrow$ II $\leftrightarrow$ III $\leftrightarrow$ I) and Axis IV is an outside vertex untouched by closure of the cycle. The compound argument is precisely that this graph structure prevents the standard response ("each problem will be solved individually") from working. Any local progress on one axis re-imports demands somewhere else; the partial-mitigation-per-axis strategy is not on track to close the cycle simultaneously, much less to do so while also producing a specification of exceeding.

Scope of the propositions. A careful reader will note that the propositions are phrased universally—"any plausible mitigation of Axis I requires substrate or paradigm changes that re-impose demands on Axis II or III," and so on. The universal reading is not the intended one. The propositions are stated against the active research program of *current* machine-learning paradigms (transformer scaling, neurosymbolic patches, agentic tool use, multimodal extension, RL from verifiers); they are not stated against a hypothetical research program that addresses multiple axes simultaneously by construction. A future maturation of the developmental-robotics tradition—one that is embodied (Axis III), sample-efficient by inductive bias (Axis II), and energy-efficient by substrate choice (Axis I) from the outset—is not refuted by the propositions, because the cross-axis dependency is a claim about how *current-paradigm* mitigations re-import

demands, not about whether some alternative paradigm could address the axes jointly by design. The operative claim (Section 1.2, "What 'infeasibility' means in this paper") already restricts the scope of the compound argument to current paradigms; the propositions inherit that scope. We accept the corresponding clarification: the compound thesis is load-bearing against the inevitability literature's actual research program (the one listed above), and is silent on whether some future qualitatively distinct program could do better. The weak inevitability claim engaged in Section 7.9 is the natural home of such cross-paradigm scenarios.

The structural claim is contestable on each proposition individually, and we have stated each with a specific refutation condition. We do not claim formal proof; we claim that the compound thesis's force depends on the propositions, and that the propositions are individually defensible against 2024–2026 evidence. A compound argument without cross-axis dependency would reduce to compound skepticism (a list of separately addressable difficulties). A compound argument with cross-axis dependency is the kind of argument the inevitability literature has not engaged, because it cannot be answered by the kind of single-axis mitigation that has answered the existing single-axis impossibility arguments. Table 1 summarizes the measured gap on each axis as of April 2026.

Table 1: The four axes of compound infeasibility, with measured 2024–2026 gaps. The table is summary; each row is developed in the corresponding subsection of Section 3 and grounded against primary sources in Section 5. The compound argument's force is that all four obtain simultaneously and that mitigating any one of them does not, on present evidence, address the others.

Axis	**Measured human reference**	**Frontier system reference**	**Approximate gap**
I. Energetic / computational	Brain ∼20 W; cortical computation ∼0.1 W; $\sim 10^{14}$–10^{16} synaptic ops/s	GPT-4 training ∼50 GWh; GPT-5 inference ∼850 MWh/day	$\sim 10^{6}$–10^{8} in per-task energy
II. Developmental / cumulative-cultural	13-year-old: $\leq 10^{8}$ words; embodied scaffolding via joint attention, teaching	Frontier LLM: 10^{12}–10^{13} training tokens; static corpus	10^{4}–10^{5} in linguistic input; not even comparable on cumulative-cultural axis
III. Causal / embodied generalization	ARC-AGI-3 panel-verified human baseline 100%; counterfactual analogies robust under surface change	Frontier LLMs <1% on ARC-AGI-3 (April 2026); sharp drop on counterfactual variants	∼100 percentage points on the current frontier benchmark
IV. Specification / target	Domain-general human cognition (informally specified)	No domain-general specification of exceeding-human general intelligence published	Conceptual gap: target is undefined

The next section makes the specification axis explicit, because it is the axis where existing literature is weakest and where the inevitability claim is most underdefined.

4 The Exceeding Problem: Why "Superhuman General Intelligence" Is Not a Coherent Target

The standard formulation of AGI requires not only matching human general intelligence but exceeding it: the inevitability claim usually frames AGI as a precursor to artificial superintelligence (ASI), and the practical motivation for AGI research is in part the expectation that such

a system will be more capable than the human researchers who built it. We argue in this section that the exceeding requirement faces an open specification problem that has not been adequately addressed in the existing literature: an ex-ante engineering specification of a domain-general exceeding target is not on offer in any existing framework, and the structural challenges to producing one (Goodhart compounding, distribution-selection ambiguity, cumulative-cultural bootstrapping) are nontrivial. The argument is conceptual rather than empirical, but it has empirical implications that we trace through to current benchmark practice.

4.1 The Specification Problem

Definition 2 (Domain-General Capability)**.** *A capability C is* domain-general *if its definition does not depend on the choice of any particular task domain T. A capability is* domain-specific *if its definition is constituted by performance on a particular T or family of T's.*

Definition 3 (Exceeding-Human Specification Problem)**.** *The Exceeding-Human Specification Problem is the problem of providing a domain-general capability C^* such that C^* exceeds the corresponding human capability C_H, in a way that is operationally measurable without reduction to any particular T.*

We pre-empt two objections that a careful reader will raise immediately.

Objection 1: this proves too much. A first reader might respond that the same argument would apply to narrow domains in which we routinely exceed human performance (chess, image classification, protein folding); since those are observably exceedable, why does the argument single out the general case? The answer is in the definitional distinction. Narrow exceeding has a domain-specific specification by construction (chess Elo, ImageNet top-1 accuracy, AlphaFold's pLDDT), and the corresponding system is super-narrow rather than super-general. The compound argument's Axis IV is not the claim that all exceeding is unspecifiable; it is the claim that the conjunction *exceeds-and-domain-general* has no specification on offer in the existing literature. The literature has produced specifications for matching (Hendrycks et al. (2025) and predecessors), and specifications for narrow exceeding (the various narrow-task benchmarks). The cell that is empty is exceeding-and-general, and the inevitability claim is precisely a claim that this cell will be filled. The empty cell is what the axis is targeted at; narrow exceedables are not within its scope.

Objection 2: this conflates feasibility with measurability. A second reader will point out that a system could in principle exceed humans without our being able to specify the target ex ante. Ex-post observation that a system has exceeded humans on some loose aggregate is not foreclosed by our argument. We accept the distinction. The axis is targeted at *ex-ante engineering specification of a domain-general exceeding target*, not at ex-post observation. The relevance of ex-ante specification is that the inevitability literature treats AGI as an engineerable artifact on a near-term timeline; engineering targets need ex-ante specifications to be optimized for, not merely ex-post diagnostic measures to be evaluated against. A research program that asserts AGI is three-to-ten years away is implicitly asserting that the exceeding target is well-defined enough to be optimized toward, and the absence of such a target in the literature is the live problem. If the inevitability claim were retreated to "AGI will eventually appear, and we will recognize it ex post," the compound argument's Axis IV is correspondingly weaker—but the inevitability claim is then also a much weaker claim, of the kind we explicitly accept (Section 7.9).

A clarification of the structural asymmetry. Human general intelligence has an *ostensive* definition: we can point at one (any cognitively typical adult) and the referent is unambiguous. Engineering a system requires a *constructive* specification: a description sufficient to evaluate whether a candidate system satisfies it. Ostensive definitions support the matching claim ("build

something that performs the way that one does"), because the target is observable. Constructive specifications are required for the exceeding claim ("build something that performs better than that one in a domain-general way"), because there is no ostensible exemplar to point at. The asymmetry is not that human cognition is well-defined and AGI is not; it is that matching can use an ostensive target while exceeding cannot, and the existing literature has not produced a constructive substitute for the missing ostensive target.

The standard claim about AGI requires a solution to the Exceeding-Human Specification Problem. The system must exceed human capability in a domain-general way; if the exceeding is only domain-specific, it is super-narrow intelligence rather than super-general intelligence, and humanity already builds super-narrow systems (chess engines, image classifiers, protein-folding predictors) without any plausible claim that their existence constitutes progress toward AGI.

The Exceeding-Human Specification Problem is not solved in the existing literature. The available specifications are all domain-specific. Composite specifications ("score above human level on the average of N benchmarks") are also domain-specific in disguise: they reduce to performance on the chosen benchmarks, and adding more benchmarks to the average does not transmute the aggregate into a domain-general measure—it merely increases the dimensionality of the domain-specific evaluation. The aggregate-of-benchmarks specification is also vulnerable to Goodhart-style overfitting: a system can saturate the chosen benchmarks (especially under the optimization pressure that frontier laboratories apply) without exhibiting the underlying capability that the benchmarks were originally taken to indicate (Goodhart, 1975; Strathern, 1997).

4.2 The Goodhart Compounding

Strathern (1997) formulated the canonical statement: "when a measure becomes a target, it ceases to be a good measure." The benchmark economy in modern AI is the textbook case: pre-training corpora are filtered, post-training data is constructed, and reward models are tuned with explicit awareness of the benchmarks on which the resulting model will be reported, with documented divergence between proxy reward and true human preference under sustained optimization pressure (Gao et al., 2022; Xu et al., 2024). The Exceeding-Human Specification Problem is acutely vulnerable: because the only available specifications are domain-specific, any program of optimizing toward AGI is a program of optimizing toward domain-specific surrogates, and each surrogate enters the Goodhart regime as soon as the optimization pressure applied to it becomes substantial.

The compounding makes this worse. Consider a hypothetical research program that optimizes for human-level-or-above performance on N diverse benchmarks. As N grows, the program tracks the inevitability claim more closely (because the aggregate covers more domains). But each individual benchmark is now subject to Goodhart pressure proportional to its visibility, and the highest-visibility benchmarks (those that drive press, customer adoption, capital allocation) experience the strongest pressure. The benchmarks at the head of the visibility distribution—the ones that anchor the AGI claim—are exactly the ones for which the surrogate–target relationship is most degraded.

The implication is not that benchmarks are useless. It is that benchmarks cannot, in the limit, substantiate a claim of domain-general superhuman capability, because the substitution of domain-specific surrogates for a domain-general target is precisely the structure Goodhart's law warns against.

4.3 Which Distribution Does Exceeding Target?

The exceeding claim is, on inspection, silent about which problem distribution it targets, and its plausibility differs starkly across the available readings. We make this explicit, because the ambiguity is doing substantial work in the inevitability literature.

Reading 1: the distribution humans actually face. Under this reading, the exceeding claim is well-posed: it asks whether a system can be built that outperforms humans on the actual mix of problems humans encounter (workplace tasks, scientific research, creative work, social reasoning). The reading is well-defined, and there is no in-principle barrier to exceeding humans on it—humans are not optimal on this distribution; biological evolution selected for fitness, not for problem-solving *per se*, and many specific subdomains (chess, image classification, protein folding) have already been exceeded by narrow systems. Under Reading 1, the question reduces to whether a *single* system can be built that exceeds humans *across the full breadth* of the distribution, simultaneously, while preserving the coherence and transfer that humans exhibit. The compound argument's first three axes apply to that question.

Reading 2: the distribution of all problems humans could in principle face. Under this reading, the exceeding claim is broader: the system must exceed humans on tasks that have not yet been encountered, including tasks whose generating distribution differs from anything in the training data. Wolpert and Macready (1997)'s No-Free-Lunch theorem is relevant here: averaged over all possible problem distributions, no algorithm outperforms any other, and an algorithm that does well outside the distribution it was trained on is not guaranteed to exist absent specific inductive biases matched to the new distribution. Reading 2 is the strong-AGI reading; it makes the exceeding claim formally hard to evaluate, because performance on distributions that have not been observed cannot be measured in advance.

Reading 3: the distribution of problems whose solutions would constitute a transformative scientific or technological advance. This is implicit in much of the inevitability literature (e.g., the framing that AGI will compress decades of scientific progress into years). Under this reading, the relevant distribution is not arbitrary but is biased toward problems whose solutions would be socially or economically valuable. The reading is more restrictive than Reading 2 but less measurable than Reading 1. The Khoja et al. timeline argument we engage in Section 7.6 effectively assumes a version of this reading.

The inevitability claim moves between these readings without committing to any of them. Reading 1 reduces the strong AGI claim to a coordination problem (combine many narrow superhuman capabilities into one coherent system, which is itself the compound-axis problem); Reading 2 invokes capabilities that cannot be measured in advance; Reading 3 begs the question of which scientific advances are achievable absent capabilities the inevitability literature has not specified. None of the three readings supports the exceeding claim as a continuous extrapolation of capability progress in the way the matching claim is supported by Hendrycks-style frameworks. The exceeding specification problem is, structurally, not the same problem as the matching specification problem.

4.4 What an Exceeding Specification Would Look Like

A genuine specification of exceeding-human general intelligence would have to:

1. Define the problem distribution explicitly (and defensibly, against the structure objection above).

2. Provide a domain-general performance measure on that distribution that does not reduce to any individual benchmark.

3. Account for the fact that human performance on the distribution is itself bootstrapped by cumulative culture, so that the comparison is meaningful: a system trained on a snapshot of cumulative culture and evaluated against humans embedded in the cumulative-cultural process is being compared against a distinct cognitive substrate.

4. Be robust to Goodhart pressure: the measure must remain meaningful under sustained optimization pressure proportional to its commercial visibility.

5. Be falsifiable: there must exist achievable observations under which the system would, demonstrably, fail to exceed humans, and these observations must be specified in advance rather than identified post-hoc.

Most of the existing literature on AGI specification gestures at these requirements without satisfying them simultaneously. The closest approximations—Humanity's Last Exam (Phan et al., 2025), ARC-AGI-2 (Chollet et al., 2025), the FrontierMath benchmark—are valuable as discriminative measures of capability progress in the regime where humans still outperform machines, but they are domain-specific in the formal sense and do not aggregate into the domain-general specification that the exceeding claim requires.

4.5 The Hendrycks Specification: A Serious Attempt at Matching, but Not Exceeding

The most serious attempt to date at the specification problem is Hendrycks et al. (2025), *A Definition of AGI*, published October 2025 by a 33-author consortium including Yoshua Bengio, Gary Marcus, Max Tegmark, Eric Schmidt, and Dawn Song. The paper defines AGI as "matching the cognitive versatility and proficiency of a well-educated adult" and operationalizes that definition by adapting the Cattell–Horn–Carroll (CHC) theory of human cognition—arguably the most empirically validated psychometric framework of the last century—into ten cognitive domains (reasoning, working memory, long-term storage and retrieval, perception, processing speed, knowledge, quantitative ability, reading and writing, auditory and visual processing). Frontier systems are scored by adapting human psychometric batteries to AI evaluation. The headline numbers: GPT-4 (2023) is 27% of the way to AGI; GPT-5 (2025) is 58%[1]. The paper diagnoses contemporary models as exhibiting a "jagged" cognitive profile—strong on knowledge tasks but with critical deficits in long-term memory storage and retrieval, suggesting that current scaling has produced uneven progress across the CHC domains rather than coherent advancement.

We engage Hendrycks et al. directly because it is, at present, the most defensible specification proposal in the literature. It satisfies, in our reading, three of our five requirements *for the matching question*: it defines a problem distribution (CHC-grounded cognitive domains), provides a partially domain-general measure (aggregated across ten subdomains rather than reduced to one benchmark), and is falsifiable for matching (the framework is published with concrete scoring methodology, and a system can be shown to fall short of the well-educated-adult target). It does not engage falsifiability for the exceeding question, because criterion 5 of Section 4.4 was articulated for the exceeding specification ("observations under which the system would, demonstrably, fail to exceed humans") and Hendrycks does not address exceeding at all. It does not satisfy the cumulative-culture requirement: CHC is a measure of human cognitive performance, but humans are embedded in the cumulative-cultural process that produced their CHC profiles, and an AI scored against the same battery is being compared against a distinct cognitive substrate. And it does not address the Goodhart-robustness requirement. The decisive observation is that the framework's own diagnosis of contemporary models is a "jagged" cognitive profile—strong on knowledge tasks but with critical deficits in long-term memory storage and retrieval. This profile is precisely the Goodhart pattern Section 4.2 predicts: aggregate scores rise (because some domains are saturated under directed optimization pressure) while other domains remain catastrophically behind, and arithmetic-mean aggregation across the ten CHC domains can mask the single-domain deficits. The pattern is therefore visible in the Hendrycks framework before any aggregation choice; it is reported by the Hendrycks authors themselves. Fourati (2025), a single recent preprint, sharpens the picture by re-aggregating the Hendrycks scores under a coherence-based scheme (geometric mean over compensability exponents); the resulting corrected score for GPT-5 is substantially below the headline 58%. Fourati has not

[1]Hendrycks et al. v1 reports 57%; v2 (post-revision) reports 58%. We adopt 58% throughout.

been independently reproduced, and we do not rest the structural argument on the specific corrected number—only on the underlying jagged-profile diagnosis, which is unambiguous in the Hendrycks paper itself. Fourati sharpens the pattern; it does not create it.

More importantly, and this is the central limitation for our argument: Hendrycks et al. is a specification of *matching*, not *exceeding*. The target is a well-educated adult, anchored to a psychometric framework whose calibration is to human populations. The framework gives no purchase on what "above-human-level general intelligence" means: "200% on the Hendrycks AGI score" is not a defined object, because the scoring methodology is bounded by human-comparable performance on each domain. CHC measures the cognitive profile of human adults; it does not extend, on its own terms, to a profile that exceeds them. The exceeding question is conceptually orthogonal to the Hendrycks framework, not addressed by it.

The honest assessment is therefore: Hendrycks et al. substantially advances the matching-specification literature; the matching specification is not the same as the exceeding specification; the exceeding specification remains, in our reading, unsolved; and even the matching specification is showing Goodhart symptoms within weeks of publication. The fourth axis of the compound argument (Section 3.4) is therefore narrowed but not retracted: matching specifications are a real and developing area, exceeding specifications are not.

4.6 Three Convergent Lines of Evidence the Specification Problem Is Live

Three lines of recent empirical work converge on the conclusion that even the matching specification is more contested than the headline scores suggest, and that all of them have direct implications for the exceeding question.

Reasoning is approximate retrieval, not generalization. Kambhampati et al. (2024) shows empirically that the apparent reasoning abilities of LLMs are dominated by exemplar-query similarity and approximate retrieval rather than systematic reasoning, and that LLMs cannot, by themselves, do planning or self-verification. The proposed remedy (LLM-Modulo frameworks combining LLMs with external symbolic verifiers) again imports neurosymbolic architecture, vindicating Marcus (2025) and supporting our Section 2.2.

Mathematical reasoning is fragile under surface-form variation. Mirzadeh et al. (2024) (Apple, ICLR 2025) introduces GSM-Symbolic, a benchmark that varies only the numerical values or surface phrasing of GSM8K problems while preserving abstract structure. Model performance drops substantially under numeric-value substitution and deteriorates further as problem complexity grows. On the GSM-NoOp variant—which appends a single irrelevant but plausible-looking clause—accuracy falls by up to ∼65% on smaller open-weight models (Phi-3-mini) and by tens of percentage points on frontier proprietary models, with the authors concluding that "current LLMs cannot perform genuine logical reasoning; they replicate reasoning steps from their training data." This is direct empirical support for the causal-grounding axis (Section 3.3), and it shows that the matching-specification's strong-reasoning entry is not as solid as the aggregate scores suggest.

Reasoning models collapse at high complexity. Apple ML Research (2025) (Apple, June 2025) studies Large Reasoning Models (LRMs) that use extended chain-of-thought, and identifies three regimes: standard models outperform LRMs on low-complexity problems, LRMs help on medium complexity, and *both collapse* on high complexity. The reasoning effort spent by LRMs increases with problem complexity up to a point and then *declines*, despite adequate token budget, and the models fail to use explicit algorithms or reason consistently across instances. The published rebuttal by Lawsen (2025) (Open Philanthropy) makes a real and partially correct critique: on the Tower of Hanoi puzzle specifically, the reported "collapse" was substantially attributable to models hitting their output token ceilings rather than to a reasoning failure (models explicitly say things like "the pattern continues, but I'll stop here to save tokens"), and Apple's evaluation pipeline judged partial-but-correct outputs as failures because it required complete enumerated move lists. When the format is changed—e.g., asking for

a recursive Lua function instead of an exhaustive move list—Claude solves 15-disk Hanoi without difficulty. We concede this critique on Tower of Hanoi specifically. The broader finding survives, however, on three grounds: (i) Apple's complexity-induced collapse pattern was reported across multiple puzzle types beyond Tower of Hanoi (river-crossing, blocks-world, checker-jumping); (ii) the GSM-Symbolic results above are independent of the Tower-of-Hanoi-specific evaluation methodology and replicate the surface-form-fragility finding through a different experimental design; and (iii) the model behavior of self-truncating mid-reasoning to save tokens is itself an empirical observation about how the models allocate reasoning effort under constraint, which the architectural axis can engage independently of whether the resulting score is interpreted as "collapse."

The convergent picture is that the matching specification, even in its best form, is being passed by systems whose underlying cognitive substrate exhibits exactly the deficits the specification was meant to test (approximate retrieval, fragile generalization, complexity-induced collapse). The headline scores rise; the underlying capability does not rise proportionately. This is the Goodhart pattern on the matching specification. The exceeding specification is even more vulnerable, because it requires a measure that exceeds human performance *coherently* rather than *on average*.

4.7 Implication: The Replication–Exceeding Asymmetry

The four axes of Section 3 apply asymmetrically to the two halves of the AGI claim. Replicating human cognition requires the system to overcome the energetic, developmental, and causal-grounding gaps. Exceeding human cognition requires, in addition, a coherent specification of what is being exceeded—and the specification problem is, we argue, conceptually separate from the replication problem.

The asymmetry has a strategic implication. An inevitability claim that relies on framing AGI as "human-level intelligence then superhuman intelligence as a continuous extrapolation" is exploiting a definitional ambiguity: the continuous extrapolation works only if the target is itself well-defined. Existing specifications work for human-level intelligence (they reduce to "performs the things humans perform, at the level humans perform them") but fail to extend to the superhuman case, because there is no human-level reference point to compare against in the regime above human performance. The continuous extrapolation is a metaphor, not a specification, and the inevitability claim has been free-riding on the metaphor.

5 Empirical Grounding: What 2024–2026 Evidence Says

The compound argument depends on quantitative claims. We summarize the evidence supporting each claim, drawn from primary sources published or updated between 2024 and 2026.

5.1 Sample Efficiency Gap

The BabyLM Challenge (Hu et al., 2024) fixes training data to either 10M or 100M words, the latter chosen to match the upper-bound estimate of cumulative linguistic input to a 13-year-old human (Hart and Risley, 1995). The 2024 challenge reports that, even after two years of community optimization, the best 100M-word submissions still trail substantially behind frontier LLMs trained on three to four orders of magnitude more data, on standard downstream evaluations including BLiMP (Warstadt et al., 2020), GLUE (Wang et al., 2018), and EWoK (EWoK, 2024). We do not claim a direct comparison to children's performance on these benchmarks (the BabyLM evaluations compare LMs against LMs, not against child psycholinguistic data); the relevant finding is that, at the data scale matched to a 13-year-old's lifetime input, current architectures do not approach the competence that the same architectures

exhibit at trillion-token scale. The disparity has not closed at a rate consistent with closure on a frontier-LLM training-run timescale.

For comparison, frontier LLMs trained on 10^{12}–10^{13} tokens (Hoffmann et al., 2022; Villalobos et al., 2024) do exhibit linguistic competence comparable to or exceeding adult humans on many measures, but only at four-to-five-orders-of-magnitude greater data input. Per linguistic token, humans appear to extract roughly four orders of magnitude more usable signal than current LLMs. The token-to-token comparison is itself a coarse measure: human linguistic exposure is embedded in ∼20 years of multimodal embodied experience and extensive cumulative-cultural scaffolding, whereas LLM training is on text alone, so the relevant cross-substrate comparison is not "tokens-to-tokens" but "tokens-to-(tokens-plus-embodied-scaffolding)." We do not claim linguistic tokens are equivalent across substrates. The relevant LLM-vs-LLM finding—which is what the BabyLM line of work measures directly—is that the same architectures, at the linguistic-input scale of a 13-year-old, do not approach the competence those architectures exhibit at trillion-token scale; the gap has not closed at a rate consistent with closure on a frontier-deployment timescale. The cross-substrate gap (LLM-vs-human) is a separate and stronger observation, but its strength as a comparison is moderated by the multimodal scaffolding caveat above.

5.2 Generalization Gap and the Benchmark-Closure Pattern

The trajectory of the ARC-AGI benchmark family (Chollet et al., 2025; ARC Prize, 2026b) is the cleanest available evidence of the causal-grounded generalization gap. ARC-AGI-2 was designed to resist memorization and to require composition of abstract operators. Through 2024 and most of 2025, frontier-model performance on ARC-AGI-2 remained far below human levels. By early 2026, the gap had closed; by April 2026, ARC-AGI-3 (an interactive successor) had immediately re-opened a 100-percentage-point gap. Table 2 summarizes the trajectory.

Table 2: ARC-AGI benchmark trajectory, 2024–2026. ARC-AGI-2 has been substantially closed by frontier reasoning models with extended thinking budgets and hybrid search/verification. ARC-AGI-3, launched March 2026, has re-opened the gap. Human-baseline conventions differ across rows: the ARC-AGI-2 individual-human baseline (∼53%) reflects a single human attempting tasks under panel-elicitation conditions; the ARC-AGI-3 panel baseline (100%) reflects multiple humans cross-checking until every task is solved. Calibrated panels approach ceiling on both benchmarks; per-task individual humans solve a substantial majority but not all. Sources: Chollet et al. (2025); ARC Prize (2025, 2026a,b); MindStudio (2026); Poetiq (2026).

Benchmark	System	Score	Approx. cost/task
ARC-AGI-2 (static abstract reasoning, launched 2025)			
ARC-AGI-2	GPT-5 (mid-2025)	9.9%	$0.73
ARC-AGI-2	Opus 4.5 Thinking (64k)	37.6%	$2.20
ARC-AGI-2	Poetiq on Gemini 3 Pro (refined)	54%	$30
ARC-AGI-2	Gemini 3 Deep Think (early 2026)	84.6%	$13.62
ARC-AGI-2	GPT-5.5 (April 2026)	85.0%	$1.87
ARC-AGI-2	Calibrated individual-human baseline	∼53%	$5 (panel)
ARC-AGI-3 (interactive, launched March 25, 2026)			
ARC-AGI-3	GPT-5.4	∼0%	N/A
ARC-AGI-3	Claude Opus 4.6	0.25%	N/A
ARC-AGI-3	Gemini 3.1 Pro	0.37%	N/A
ARC-AGI-3	Arcgentica (Opus 4.6 + program-synthesis harness, Symbolica)	36.08%	$1,005 (full run)
ARC-AGI-3	StochasticGoose (non-LLM: CNN + RL, Tufa Labs)	12.58%	N/A
ARC-AGI-3	Blind Squirrel (non-LLM: state-graph search)	6.71%	N/A
ARC-AGI-3	Human baseline (panel)	100%	N/A

The trajectory has four observations relevant to the compound argument. First, ARC-AGI-2 was closed not by scaling but by the addition of explicit reasoning machinery: extended chain-of-thought, search over candidate solutions, program-synthesis components, structured verification of intermediate steps. The closure is direct empirical support for Marcus (2022, 2025). Pure scaling did not close the gap; hybrid neurosymbolic methods did. Second, the closure required, in compute terms, hundreds to thousands of times the per-task energy a human uses on the same problem, even granting the fast end of cost estimates. The gap closed in score; it did not close in efficiency. Third, ARC-AGI-3 was launched almost immediately after ARC-AGI-2 was closed, and the immediate re-opening of a 100-point gap suggests that the closure of ARC-AGI-2 did not transfer to a new format that requires interactive rule discovery. The closure was specific to the static-grid format, not to the underlying capability of causal-grounded generalization. Fourth, the systems that close substantial fractions of the gap on ARC-AGI-3 fall into two categories, both of which import architectural machinery the default LLM configurations lack: bespoke harnesses on top of frontier LLMs (Symbolica's Arcgentica, an Opus-4.6-based program-synthesis orchestrator at 36.08%) and purpose-built non-LLM agents (StochasticGoose's CNN with reinforcement learning at 12.58%, Blind Squirrel's state-graph search at 6.71%). Each outperforms default-configuration frontier LLMs by one to two orders of magnitude. The Arcgentica result is particularly informative: the same Opus 4.6 model goes from 0.25% in default deployment to 36.08% inside an orchestrator, at one-tenth the cost—direct evidence that the capability gain is in architectural orchestration, not in the model's parameters or training data, and that the architectural primitives in default LLM-chat use are not even directionally well-suited to the format.

The pattern across the trajectory is the data: each generation of benchmark designed to probe causal-grounded generalization is eventually closed by importing architectural machinery, the closure is compute-intensive and format-specific, and the next benchmark exposes the next layer of the gap. The compound argument predicts exactly this pattern, because closing the underlying constraint (sample-efficient causal learning grounded in embodied interaction) requires research programs that the closure mechanisms do not constitute.

5.3 Data Exhaustion

Villalobos et al. (2024) (Epoch AI) estimate the effective stock of quality-and-repetition-adjusted human-generated public text at ∼300 trillion tokens, with an 80% confidence interval that this stock will be fully utilized by frontier training runs at some point between 2026 and 2032. The interval depends on how aggressively models are overtrained relative to compute-optimal: at standard Chinchilla-optimal training (Hoffmann et al., 2022), the stock supports models trained with up to $\sim 5 \times 10^{28}$ FLOP, a level expected to be reached by approximately 2028. At the overtraining levels used by some frontier labs (5× or higher), the stock could be exhausted as early as 2027.

The implication is that the dominant input to the scaling regime—human-written text on the open web—has a quantifiable upper bound, and the upper bound is being approached on the timescale on which frontier labs have publicly committed to AGI. The standard responses (synthetic data, agentic data generation, multimodal extension) extend the input surface but do not extend the underlying signal: synthetic data, by construction, contains only the patterns implicit in the model that generated it, and the quality of synthetic data eventually bottoms out on the quality of the original human-generated training distribution (Shumailov et al., 2024; Gerstgrasser et al., 2024).

5.4 Researcher Consensus

The AAAI 2025 Presidential Panel on the Future of AI Research (AAAI, 2025) surveyed 475 AI researchers, of whom 76% rated it "unlikely" or "very unlikely" that scaling current approaches

would deliver AGI. The panel itself, composed of 25 senior researchers (with 15 additional contributors) across infrastructure, methods, and societal impact, concludes that progress toward more general intelligence requires moving beyond big-data scaling and engaging the architectural, embodied, and social-learning dimensions discussed above.

We cite the survey not as proof of infeasibility—majority opinion is not proof—but as evidence that the inevitability narrative is not a research consensus. The narrative is concentrated in industry communications and in a smaller subset of researchers; the broader research community, when surveyed, is substantially more skeptical. The compound argument is, in this sense, an articulation of a position closer to the median researcher view than to the median industry communication.

5.5 Energy and Compute Constraints

Data center electricity consumption was 5.2% of U.S. total in 2025 (IEA, 2025), with International Energy Agency projections of 8–10% by 2030. The growth is driven primarily by inference at scale rather than training: Epoch AI (2025) estimate that GPT-5-class inference consumes on the order of 850 MWh per day, with serving costs over a 90-day period substantially exceeding the corresponding training costs. The shift from training-bound to inference-bound compute has implications for the scaling argument: even if training compute continued to grow at recent rates, inference compute is the dominant binding constraint on deployed capability, and inference compute scales with deployment surface rather than with model size alone.

The energetic comparison to the human brain is unfavorable. Frontier inference per task consumes on the order of 10^4–10^5 J at typical reasoning-token volumes; the brain consumes on the order of 10^1–10^2 J per equivalently complex reasoning task. The ratio is on the order of 10^2–10^4 per task, and the gap widens when the comparison is made for the kind of tasks (open-ended reasoning, novel-domain transfer) on which the brain is most dominant.

5.6 Summary

The empirical evidence does not, on its own, prove infeasibility—no empirical evidence can prove a claim of formal infeasibility. It does establish that the four axes of the compound argument are quantitatively grounded: the sample-efficiency gap is real and measured in orders of magnitude; the generalization gap is real and measured by benchmarks specifically designed for the purpose; the data-exhaustion bound is documented and on the timescale of the AGI claim; the researcher consensus is documented; the energetic gap is documented. The compound argument's force is that all of these gaps obtain simultaneously, and that the partial mitigations applied to any one of them are not, on the available evidence, closing the others.

6 Falsifiability: What Would Refute the Compound Argument?

A serious infeasibility claim should specify, in advance, what evidence would refute it. We propose three refutation conditions, each independently sufficient. The conditions are stated quantitatively where possible. If any of them is met by 2030, the compound argument should be retracted.

Refutation Condition 1 (Sample-Efficiency Threshold)**.** *A model architecture is demonstrated to achieve adult-human-comparable linguistic competence on a defensible aggregate of measures (BLiMP, GLUE, EWoK, BabyLM downstream) when trained on no more than* 10^9 *tokens of natural language input, with no auxiliary corpus and no auxiliary embodied or multimodal supervision, and replicates this result across at least three independent reproductions.*

The threshold is one order of magnitude above the BabyLM 100M-word constraint, which would represent a substantial closure of the sample-efficiency gap without requiring exact match

to the human input curve. The auxiliary-supervision exclusion aligns this condition with Proposition 2 (Developmental-Axis Dependency), which argues that closing Axis II without auxiliary supervision is the load-bearing case; the supervision-permitted case is consistent with our argument and would not refute it. The condition refutes the developmental axis: closing the sample-efficiency gap on a text-only training pipeline would demonstrate that the developmental scaffolding can, in some form, be replaced by an alternative that current architectures do not require.

Refutation Condition 2 (Transfer-Without-Retraining Threshold)**.** *A model achieves human-comparable performance ($\geq$ 80% of human baseline) on ARC-AGI-2 (or a successor benchmark with equivalent contamination resistance and out-of-distribution structure), at a per-task compute cost not exceeding $10^3\times$ human task time when both are measured in joules, without task-specific fine-tuning.*

The threshold is, again, conservative: a 10^3 energy ratio still leaves humans dramatically more efficient, but it represents the closure of the most measurable causal-grounded generalization gap. The condition refutes the causal axis.

Refutation Condition 3 (Specification-and-Realization Threshold)**.** *A documented domain-general specification of exceeding-human general intelligence is published that satisfies the five criteria of Section 4.4, accepted by a representative cross-section of the AI research community (operationally: adopted as the headline metric in three or more independent benchmarks within 24 months of publication, or endorsed in a community consensus document of comparable standing to AAAI (2025) or Hendrycks et al. (2025)), and a system is empirically demonstrated to satisfy it under at least two independent reproductions.*

The third condition refutes the specification axis. It is the highest-bar condition because it requires both a conceptual achievement (the production of a domain-general specification of exceeding, which the existing literature has not produced) and an empirical one (a system that satisfies it). The condition is in principle satisfiable; Proposition 4 (Specification-Axis Independence) does not foreclose the possibility, only asserts that mitigations of the empirical axes do not, on their own, produce such a specification. We include it because the compound argument depends on this conceptual gap; without a specification, the inevitability claim's framing of AGI as engineerable on a single-digit-year timeline is targeting an undefined object.

Claim 2 (Burden of Proof)**.** *Until at least one of Refutation Conditions 1 to 3 is met, the burden of proof for the inevitability claim sits with its proponents. The default position should be that AGI as defined is not feasible on the path described by current paradigms, and that capability progress within current paradigms produces useful AI systems but not progress toward AGI specifically.*

We do not claim novelty for the principle that an extraordinary claim requires extraordinary evidence (Sagan, 1980). We claim that the inevitability literature has not, to date, provided the evidence, and that the rhetoric of imminence has been allowed to substitute for the evidentiary structure that would, in any other engineering domain, be required for a project of this stated scope.

7 Counterarguments

We engage the strongest counterarguments we have found in the literature.

7.1 "Each Gap Will Be Closed Individually"

The most common counterargument is that each of the four gaps is, individually, the subject of an active research program, and that there is no in-principle reason any of them cannot be closed. We grant the premise. The compound argument's response is that the research programs targeting each gap are not, in general, the same research program, and that the timelines for the four programs are not synchronized. Closing the energetic gap requires neuromorphic hardware. Closing the developmental gap requires sample-efficient learning architectures fundamentally distinct from transformers. Closing the causal gap requires embodied agentic systems with intervention surfaces orders of magnitude broader than current tool-use APIs. Closing the specification gap requires conceptual work that is not, structurally, an engineering problem at all.

The response that "each problem is being worked on" is not a response to the compound argument; it is a list of partial responses to each axis individually. The compound argument requires that all four gaps be closed approximately simultaneously, because closing one gap re-imports the others. The cross-axis dependency in Section 3.5 is the structural form of the claim: any plausible mitigation of one axis demonstrably re-imports demands on another (a neuromorphic system that solves the energetic gap still needs developmental scaffolding—Proposition 1; a sample-efficient architecture still needs causal grounding—Proposition 2; an embodied agent still needs an exceeding specification—Proposition 4). The synchronization requirement is, on the available evidence, not on track to be met.

7.2 "The Brain Is Just a Computer; What It Does, Silicon Can Do"

The Church–Turing thesis establishes that any computable function can in principle be computed by any universal computational system. The thesis is sometimes invoked against any infeasibility claim about AGI. The response is that the thesis is correct but does not deliver what the inevitability claim requires. "In principle computable" is not the same as "feasible to compute on a timescale and resource budget compatible with the inevitability claim." The compound argument is not that AGI violates Church–Turing; it is that the resource and timeline requirements are several orders of magnitude beyond what is currently plausible, and that the partial mitigations being applied do not aggregate.

7.3 "Scaling Laws Are Predictive: Why Expect the Loss Curves to Stop Falling?"

A pro-scaling counterargument distinct from the emergence claim observes that compute-and-data scaling laws (Hoffmann et al., 2022) have been remarkably predictive of pretraining loss across multiple orders of magnitude, that the cross-entropy curves continue to fall smoothly with additional compute, and that this predictiveness gives the inevitability claim an empirical foundation independent of any specific capability extrapolation. The compound argument's response has three parts.

First, predictive cross-entropy curves measure performance on the pretraining objective (next-token prediction on a held-out distribution drawn from the same source as the training data), not on the capabilities the AGI claim targets. Loss continues to fall in domains where it has been falling; this is a different claim than that capabilities continue to scale in domains the loss curves do not directly measure. Schaeffer et al. (2023) and the GSM-Symbolic results above (Section 4.6) document exactly this gap: smooth loss-curve progress on pretraining objectives, jagged or absent progress on causal-grounded generalization targets.

Second, the scaling-law program is itself decelerating in the regime that matters for the AGI claim. Villalobos et al. (2024) document that the data side of the Chinchilla equation faces exhaustion on a 2026–2032 timescale; AAAI (2025) document that 76% of surveyed researchers

no longer believe further scaling will deliver AGI. The smoothness of the loss curve is not in dispute; the question is whether further compute on a scaling regime approaching its data ceiling continues to deliver capability gains commensurate with the inevitability claim's required pace.

Third, the smoothness argument is symmetrical with respect to the conclusion: if the loss curves fall smoothly while the capability gaps documented in Axes I–III persist, that is itself evidence that the loss-curve metric and the capability target are dissociable. The compound argument is precisely about that dissociation.

The emergence argument holds that frontier capabilities have repeatedly appeared at scales beyond which they were predicted, and that further scaling will continue to produce qualitatively new capabilities. Recent work (Schaeffer et al., 2023) has substantially undermined the strong form of the emergence claim, showing that many supposedly emergent capabilities are artifacts of discontinuous evaluation metrics and disappear under continuous metrics. We grant the weaker form of emergence: capability progress within the in-distribution regime is real and has been faster than many predicted. The compound argument concerns capabilities outside that regime: causal-grounded generalization, sample-efficient learning, domain-general optimization. Emergence has not been demonstrated for these capabilities; the ARC-AGI-2 trajectory is the cleanest case, and frontier model gains there have been incremental and have required architectural change rather than pure scaling.

7.4 "Human Cognition Is Just Statistical Pattern Matching, Too"

Some defenders of the scaling path argue that human cognition is itself a statistical pattern-matching system, and that the differences between human and LLM cognition are quantitative rather than qualitative. The argument has support from connectionist accounts of cognition (Rumelhart and McClelland, 1986) and is not without merit. The compound argument's response is that even if human cognition is, in some abstract sense, statistical pattern-matching, the specific statistics it does match—grounded in causal interaction, scaffolded by cumulative culture, run on a 20 W energy budget—are not the statistics current LLMs match. The argument that "humans are also statistical learners" does not establish that current systems are on the path to human-level statistical learning; it establishes a bar at which the comparison becomes meaningful, and the empirical gap to that bar is what the four axes measure.

7.5 "This Argument Could Have Been Made Against the Wright Brothers"

The historical analogy argument observes that other transformative technologies (heavier-than-air flight, the internet, deep learning itself) were predicted impossible shortly before they were realized, and concludes that infeasibility arguments have a poor track record. The compound argument's response is, first, that the cited cases are domains in which the negative case was made on architectural rather than physical grounds: the Wright Brothers solved an engineering problem within known physics, not a problem that violated physical constraints. The compound argument is grounded in measured constraints (energy, sample efficiency, data exhaustion, generalization gaps), not in architectural intuition about what current systems lack. Second, the analogy works in both directions: there are far more historical predictions of imminent transformative technology that did not materialize on the predicted timescale (cold fusion, full self-driving, strong AI in the 1960s, expert systems in the 1980s) than there are predictions of impossibility that were overturned. The base rate does not, in fact, favor the inevitability claim.

7.6 "The Hendrycks/Khoja Timeline Says AGI Is Three Years Away"

The strongest live timeline argument from inside the AGI-definition literature is Khoja and Hiscott (2025) (Hendrycks-aligned, AI Frontiers, October 2025). The argument: applying the Hendrycks AGI definition, GPT-5 is at 58%; extrapolating recent capability progress, there is a

∼50% probability that frontier systems will reach >95% on the Hendrycks score by end-2028 and ∼80% probability by end-2030. The single named bottleneck is continual learning (long-term memory storage and retrieval); the rest, the argument holds, is "business-as-usual engineering." Frontier-laboratory leadership has made compatible statements. Sam Altman writes that "we are now confident we know how to build AGI as we have traditionally understood it" and that "in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies" (Altman, 2025). Dario Amodei writes that "powerful AI" (his preferred term) "could come as early as 2026, though there are also ways it could take much longer" (Amodei, 2024). Demis Hassabis estimates that AGI "will start coming to the fore" over "the next five to ten years," assigning ∼50% probability to that window (Hassabis, 2025).

Three responses. First, the timeline argument depends entirely on the Hendrycks score being a faithful proxy for AGI progress, and the framework's own "jagged-profile" diagnosis (Section 4.5) is the Goodhart compounding pattern in unambiguous form: the headline 58% aggregates over single-domain deficits that, under any coherence-aware aggregation, drag the corrected score substantially below 58%. The exact corrected number is contested (Fourati (2025) is a single preprint and we do not rest the argument on it), but the structural vulnerability is independent of the specific value. The timeline argument's linear extrapolation from 27% (GPT-4) to 58% (GPT-5) to a 2028 forecast assumes the 58% is on a smooth capability curve; the jagged-profile pattern says it is not.

Second, the argument identifies continual learning as the one needed breakthrough. We agree it is one needed breakthrough, and we agree it is not addressed by scaling current architectures. We disagree that it is the *only* needed breakthrough. The compound argument identifies four axes; continual learning sits at the intersection of Axis II (developmental scaffolding) and Axis III (causal-grounded generalization). Closing that intersection is necessary but not sufficient; Axis I (energetic) and Axis IV (specification of exceeding) remain open even after continual learning is solved. The Khoja timeline implicitly assumes that the other axes will be addressed as side effects of continual learning, which is a stronger claim than the timeline argument explicitly defends.

Third, the linear-extrapolation methodology used to convert "GPT-4 = 27%, GPT-5 = 58%" into a 2028 forecast assumes the underlying capability curve is linear and that the bottleneck remains the one currently identified. Both assumptions are contested by the empirical evidence in Section 5: the Hendrycks-domain progress is jagged rather than smooth, scaling laws are decelerating, and the benchmark-closure pattern (Section 5.2) shows that closure on one benchmark does not transfer to the next benchmark in the sequence.

The honest summary is that the timeline argument and the compound argument are not actually incompatible if the timeline is read carefully: Khoja predicts a ∼50% probability of reaching the Hendrycks >95% threshold by 2028; we predict that reaching that threshold (under the headline-arithmetic-mean aggregation) is consistent with not having addressed Axes I, III, or IV (and possibly II). The disagreement is about what the threshold measures, not about whether the threshold can be reached. Reaching 95% on the Hendrycks framework under arithmetic-mean aggregation is consistent with not having addressed the energetic, causal, or specification axes—and the Fourati coherence-corrected reading suggests that this pattern (high arithmetic-mean score, persistent domain-specific deficits) is already visible in current systems. The timeline argument's predictive power therefore depends on a measurement convention the compound argument explicitly contests.

7.7 "Synthetic Data and RL from Verifiers Solve the Data-Exhaustion Problem"

A common counterargument to the data-exhaustion axis (Section 5.3) is that DeepSeek R1, OpenAI's o-series, and the broader 2024–2026 wave of reinforcement-learning-from-verifier setups generate enormous quantities of useful synthetic reasoning traces, and these traces are not

collapsing in the Shumailov et al. (2024) sense. The argument concludes that data exhaustion is not, in fact, a binding constraint.

The response has three parts. First, the synthetic-trace argument is strongest in domains with cheap and accurate verifiers: mathematics (verify by computation), code (verify by execution), formal logic (verify by proof checker). It is weakest in exactly the domains the Hendrycks framework identifies as deficient: long-term memory, perceptual grounding, social and pragmatic reasoning, novel-domain transfer. RL from verifiers is a method for amplifying capability in domains where ground truth is computationally cheap; it is not a general method for replacing human-generated data with self-generated data of equivalent diversity.

Second, the verifiers themselves remain bottlenecked on human-generated specifications. A code-execution verifier checks whether code passes tests, but the tests are written by humans (or by models trained on human-written tests). A mathematical-proof verifier checks formal correctness, but the formal definitions and theorem statements come from human mathematics. The synthetic-trace approach moves the data bottleneck up one level—from training corpus to verifier specification—rather than eliminating it.

Third, the empirical evidence on whether RLVR-driven training generalizes outside the verifier-supervised distribution is contested but not negative, and the compound argument's response should engage this honestly. The 2024–2026 wave of reasoning models trained with mathematics-and-code verifiers (DeepSeek R1, OpenAI's o-series, and derivatives) does exhibit measurable improvements on benchmarks outside the verifier-supervised distribution—reading comprehension, multi-step reasoning in non-formal domains, aggregate scores on broad evaluation suites—relative to base models trained without RLVR. The transfer is real, and a defender of the synthetic-trace approach is correct to point at it. Our response is about magnitude rather than existence: frontier RLVR-trained models continue to score near-zero on ARC-AGI-3 (Sections 3.3 and 5.2), continue to fail GSM-Symbolic surface-form variants by tens of percentage points on the same systems whose RLVR pipelines deliver the cross-domain gains (Section 4.6), and continue to exhibit the Hendrycks "jagged-profile" deficits in long-term memory and adaptive reasoning (Section 4.5). What the RLVR transfer demonstrates is partial amplification of in-distribution-adjacent reasoning—a meaningful capability gain that the compound argument does not deny—not closure of the causal-grounded generalization gap that Axis III specifies. The synthetic-trace approach is, on the available evidence, a real capability-amplification mechanism with a ceiling on out-of-distribution transfer that is empirically constrained: consistent with substantial RLVR-driven gains, inconsistent with those gains aggregating into closure of the four axes. Gerstgrasser et al. (2024) show that mixing synthetic with original data avoids collapse in some regimes; Shumailov et al. (2024) show collapse in others; the long-term-drift question under recursive synthetic training remains live, and the compound argument does not need it resolved against the inevitability claim to do its work.

7.8 "Embodied Foundation Models Will Address the Causal-Grounding Gap"

A live counterargument to the causal-grounding axis (Section 3.3) is that the 2024–2026 wave of embodied foundation models—Genie 2's interactive 3D world generation (Parker-Holder et al., 2024), OpenVLA's vision-language-action model (Kim et al., 2024), π_0's flow-matching dexterous manipulation across seven robot platforms (Black et al., 2024), RT-2 (Brohan et al., 2023) and its successors—introduces exactly the intervention surface that text-only training lacks. The argument concludes that the causal-grounding axis is being addressed by extending the modality of training rather than by changing paradigm.

We grant the trend. The post-2024 embodied wave is genuinely the most plausible direction for closing the causal-grounding gap, and we have said so (Section 3.3). Three caveats limit the force of the counterargument.

First, the breadth of the intervention surface remains far short of what supports human causal learning. OpenVLA is trained on $\sim 10^6$ robot manipulation trajectories from the Open-X

Embodiment dataset (Kim et al., 2024); π_0 trains on data from 7 robot platforms and 68 unique tasks (Black et al., 2024). A human child accumulates more diverse intervention experience in the first year of life (every push, drop, taste, throw, balance, reach) than any current embodied dataset captures in aggregate. The dataset-scale gap is comparable to the BabyLM/frontier-LLM gap on the developmental axis, and is not closing on a frontier-deployment timescale.

Second, embodied training does not, on the available evidence, transfer cleanly to non-embodied reasoning. A model trained on robotic manipulation has acquired causal models of object-pushing; whether those models transfer to reasoning about, say, social interactions or counterfactual scenarios in language is an open empirical question, and the early evidence from VLA evaluation suites suggests transfer is highly task-specific.

Third, the energetic axis (Section 3.1) becomes more rather than less binding under embodied training. Robot intervention is more expensive than text scraping, by orders of magnitude per useful sample. Genie 2 mitigates this by generating interactive worlds in simulation, but simulation does not introduce causal intervention in the world; it introduces intervention in a model of the world, which is a weaker training signal. The post-2024 embodied wave is therefore the right direction for closing Axis III, but it imports tighter constraints on Axes I and II than the original architectures faced. The compound argument predicts exactly this pattern.

7.9 "Even If Slow, We Will Get There Eventually"

The weakest form of the inevitability claim is that AGI may not be on the timeline the industry suggests, but it will be reached eventually. The compound argument is compatible with a sufficiently weak version of this claim. If "eventually" is allowed to extend beyond the timescale on which the four axes are addressed by qualitatively new research programs (not merely extensions of current paradigms), then the argument is not that AGI is impossible, but that AGI is not on the path the field is currently on. We accept that framing. The strong inevitability claim—that AGI is imminent on a single-digit-year timescale and is a continuous extrapolation of current methods—is the claim the compound argument is designed to refute. The weak inevitability claim—that AGI is achievable in some indefinite future given research programs distinct from current ones—is not refuted by our argument and is not the claim the inevitability literature is making.

8 Limitations

The compound argument is qualitative-with-empirical-grounding rather than formally proved. We list, explicitly, the limitations of the argument as presented.

The argument is not a formal impossibility proof. van Rooij et al. (2024) present a complexity-theoretic theorem; we present a structural argument across four axes, each grounded in measured quantities and cited primary sources. The structural argument has the advantage of avoiding the Guerzhoy (2024) critique of formal proofs but the disadvantage that no individual axis admits a clean theorem. The compound argument's force is in the conjunction; the conjunction is not formally provable. We have stated this openly. Readers who require a formal infeasibility proof for the inevitability claim should treat this paper as a complement to, not a substitute for, the formal-impossibility literature.

Empirical claims are subject to revision. The benchmark trajectory is, by definition, a moving target. ARC-AGI-2 was the live frontier-generalization benchmark for most of 2025; by April 2026 it had been substantially closed, and ARC-AGI-3 had become the live benchmark. The compound argument's structure (each closure prompts the next benchmark, each closure is compute-intensive and format-specific) is robust to these movements; the specific numbers in

Table 2 are not. Readers reading this paper after ARC-AGI-3 has been closed should look for the corresponding successor benchmark and assess whether the closure pattern has continued or has been broken.

Order-of-magnitude estimates. Several quantities reported in this paper (per-task inference energy, brain energy partition between computation and communication, total synaptic operations per second) are order-of-magnitude estimates from the literature, not point measurements. The compound argument is robust to the specific numbers within their plausible ranges; an order-of-magnitude shift in any single number would not change the structure of the argument. Readers who require tighter precision should consult the cited primary sources, where the underlying methodologies and uncertainties are documented.

Definition of AGI. We have used a single definition of AGI throughout: a system matching and exceeding human general intelligence across the full range of intellectual tasks. The definition is consistent with the inevitability literature (Shah et al., 2025; Fortune, 2025) and with the dominant industry usage. It is not the only possible definition. Some authors restrict AGI to "human-comparable performance on a sufficiently broad benchmark suite," which removes the exceeding requirement and reduces the compound argument to its first three axes; others extend AGI to include consciousness or moral status, which broadens it beyond the scope of this paper. The compound argument is targeted at the dominant definition; we have not attempted to argue against or for alternative definitions.

Conflict of interest. The author is the founder of Rōmy, a company that builds vertical-AI products using frontier-LLM infrastructure. This paper was written independently of the company; it received no Rōmy funding or company resources, and reflects the author's personal research rather than the company's position. The argument that AGI is not on the path of current paradigms is consistent with the commercial position of vertical-AI companies generally (which sell AI systems, not AGI), and a reader should weigh the argument with that affiliation in mind. A less direct incentive is also worth flagging: the case that AGI is not imminent on current paradigms is congenial to founders of vertical-AI companies more broadly, because it justifies a market structure in which vertical specialization, rather than horizontal AGI sweep, is the dominant mode of value capture; a founder making such an argument is not making it from a position of disinterest. We flag this second-order incentive not to retract the argument but to offer the reader a fuller picture for calibration. No claim in this paper depends on Rōmy's products or commercial positioning, and the argument's structure is independent of any particular vendor's success or failure.

What we are not arguing. We are not arguing that AI is not transformative, that frontier laboratories are not doing valuable research, that current LLMs are without utility, or that the trajectory of capability progress over the last decade has been illusory. The compound argument is narrowly about the specific proposition labeled "AGI" in the inevitability literature: a system matching and exceeding human general intelligence across domains. Useful, valuable, even transformative AI systems are entirely compatible with the infeasibility of AGI as defined.

9 Implications and Conclusion

9.1 Implications for AI Research

If the compound infeasibility argument is correct, the implications for AI research are substantial but not dramatic. The argument is fully compatible with continued capability progress within current paradigms. The systems that will be built in the next decade—larger context windows,

more sophisticated reasoning chains, broader tool use, multimodal integration, improved sample efficiency, better alignment—will be valuable, will be commercially substantial, and will substantially expand what AI can usefully do. None of this constitutes progress toward AGI as defined; all of it constitutes progress toward better narrow and semi-general AI. The two trajectories are not the same trajectory.

The implication for research strategy is that programs that frame themselves as "working toward AGI" should be evaluated on their progress toward the four axes, not on their progress toward larger models or longer reasoning chains. A research program that addresses sample efficiency by an order of magnitude is genuinely on the AGI path; a program that scales an existing architecture is not. A program that introduces causal intervention into the training loop is genuinely on the AGI path; a program that adds tool use to inference is not. The compound argument provides a checklist for evaluating which research programs are addressing the structural gaps and which are accumulating capability within the existing paradigm.

9.2 Implications for AI Policy

The policy implications follow directly. AI policy that is structured around an imminent AGI transition is structured around an event the compound argument suggests will not happen on the assumed timescale. The risks of current AI systems—misuse, unemployment-driven economic disruption, concentration of capability in a small number of frontier laboratories, security and information-integrity concerns—are substantial and present. The policy framework should engage them on their own terms, not as precursors to a transition that the evidence does not support.

The argument is also compatible with the claim that capability concentration in frontier laboratories is itself a substantial risk, independent of the AGI question. The compound argument addresses what frontier capabilities are not (a path to AGI); it does not address what they are (substantially capable systems with substantial influence over information environments, economic activity, and decision-making). Policy that conflates the two—treating "superhuman general capability" as the relevant concept when the actual concept is "narrowly-superhuman, broadly-capable, commercially concentrated systems"—will mismatch its instruments to its targets.

9.3 Implications for Public Discourse

The compound argument is, finally, an argument that the public discourse around AGI has been calibrated to a claim that the evidence does not support. The discourse has been organized around imminence and inevitability; the compound argument is that the imminence is not established and the inevitability has no formal foundation. The appropriate epistemic posture for the public is, in our view, neither "AGI is coming soon, prepare accordingly" nor "AGI is impossible, ignore it," but "the systems being built are transformative within their actual domains; the AGI claim that frames their development is, on inspection, weakly supported, and the burden of proof for that claim sits with its proponents."

9.4 Conclusion

The compound infeasibility argument holds that AGI as defined—a system matching and exceeding human general intelligence across domains—is not feasible on the path described by current machine-learning paradigms, because four reinforcing constraints, none individually fatal, jointly preclude any plausible path on the operative timeline: the energetic and computational economy of human cognition, the developmental and cumulative-cultural scaffolding that human cognition runs on, the causal-embodied grounding that supports human generalization, and the absence in the existing literature of a domain-general specification of *exceeding* that target. We

use "infeasibility" in the scoped sense set out in the operative-claim paragraph of Section 1.2: not in-principle impossibility, but infeasibility-on-the-path-of-current-paradigms, on the timeline the inevitability literature describes.

The argument is structured to avoid the vulnerability that has limited the existing impossibility literature. It does not depend on adversarial-distribution assumptions; it does not deny inductive biases; it does not rest on a single architectural feature. The compound thesis's force depends on the cross-axis dependency we develop in Section 3.5: any plausible mitigation of one axis re-imports demands on at least one other axis, so the rebuttals to single-axis arguments do not aggregate into a rebuttal of the conjunction. Each axis is grounded in measured 2024–2026 quantities; each is associated with a specific refutation condition under which the argument should be retracted.

We do not claim that the systems being built today are without value, that the field is on a wrong track in any general sense, or that progress within current paradigms is illusory. We claim that the specific proposition of AGI is not a scaling problem, not a near-term engineering target, and not a continuous extrapolation of capability progress. The path to AGI, if there is one, runs through sample-efficient architectures, embodied causal learning, dynamic embedding in cumulative-cultural processes, and conceptual work on the specification of what exceeding human general intelligence would mean. None of these is an extension of the dominant paradigm. All of them are research programs whose connection to the dominant paradigm is, at most, a shared label.

Until the refutation conditions are met, the burden of proof sits with the inevitability claim. The compound argument does not need to be proved beyond reasonable doubt; the inevitability claim does, and it has not been.

Acknowledgments

The author thanks the Rōmy team for ongoing discussions about the boundary between current AI capability and the AGI claim, and acknowledges intellectual debts to the cognitive-science literature on cumulative culture (Tomasello, Henrich, Heyes), to the embodied-cognition tradition (Lakoff, Clark, Chemero), to the formal-impossibility literature (van Rooij, Guest, Adolfi), and to the empirical-skepticism literature (Marcus, Chollet, Mitchell). The compound structure is the author's; the constraints it composes are not.

Use of AI assistance. The author developed and refined the argument through iterative collaboration with Claude (Anthropic), accessed via Claude Code. Claude assisted with literature search and synthesis, proposed structural framings that the author evaluated and adapted, drafted and revised prose, and cross-checked citations against primary sources. All editorial decisions, the choice of thesis, the engagement with opposing literature, and the final claims are the author's responsibility. The author has reviewed all empirical claims against the cited primary sources.

References

AAAI Presidential Panel. Future of AI research: Report of the AAAI 2025 presidential panel. *Association for the Advancement of Artificial Intelligence*, March 2025. https://aaai.org/about-aaai/presidential-panel-on-the-future-of-ai-research/.

ARC Prize. ARC Prize 2025 results and analysis. https://arcprize.org/blog/arc-prize-2025-results-analysis, December 2025.

ARC Prize. ARC-AGI leaderboard. https://arcprize.org/leaderboard, April 2026.

ARC Prize. Announcing ARC-AGI-3: An interactive benchmark for AI agents. https://arcprize.org/blog/arc-agi-3-launch, March 2026.

Altman, S. Reflections. *Sam Altman's Blog*, January 2025. https://blog.samaltman.com/reflections.

Amodei, D. Machines of loving grace: How AI could transform the world for the better. *darioamodei.com*, October 2024. https://www.darioamodei.com/essay/machines-of-loving-grace.

Shojaee, P., Mirzadeh, S. I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Apple Machine Learning Research*, June 2025. https://machinelearning.apple.com/research/illusion-of-thinking.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541, 2009.

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., and Wilcox, E. G. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*, December 2024.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT 2021*, pages 610–623, 2021.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, October 2024.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Chemero, A. *Radical Embodied Cognitive Science*. MIT Press, 2009.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, November 2019.

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Clark, A. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, 2008.

Csibra, G. and Gergely, G. Natural pedagogy. *Trends in Cognitive Sciences*, 13(4):148–153, 2009.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., et al. An approach to technical AGI safety and security. *arXiv preprint arXiv:2504.01849*, 2025. Google DeepMind technical report.

Dennett, D. C. Cognitive wheels: The frame problem of AI. In Hookway, C., editor, *Minds, Machines and Evolution*, pages 129–151. Cambridge University Press, 1984.

Parker-Holder, J., Ball, P., Bruce, J., Dasagi, V., Holsheimer, K., Kaplanis, C., Moufarek, A., Scully, G., Shar, J., Shi, J., Spencer, S., Yung, J., Dennis, M., Kenjeyev, S., Long, S., Mnih, V., Chan, H., Gazeau, M., Li, B., Pardo, F., Wang, L., Zhang, L., Besse, F., Harley, T., Mitenkova, A., Wang, J., Clune, J., Hassabis, D., Hadsell, R., Bolton, A., Singh, S., and Rocktäschel, T. Genie 2: A large-scale foundation world model. *Google DeepMind Blog*, December 4, 2024. https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/.

Drachman, D. A. Do we have brain to spare? *Neurology*, 64(12):2004–2005, 2005.

You, J. How much energy does ChatGPT use? *Epoch AI Gradient Updates*, 2025. https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use.

Ivanova, A., Sathe, A., Lipkin, B., Kumar, U., Radkani, S., Clark, T. H., et al. Elements of world knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*, 2024.

Goldman, D. Google DeepMind 145-page paper predicts AGI matching top human skills could arrive by 2030. *Fortune*, April 2025. https://fortune.com/2025/04/04/google-deeepmind-agi-ai-2030-risk-destroy-humanity/.

Fourati, F. A coherence-based measure of AGI. *arXiv preprint arXiv:2510.20784*, October 2025.

Frith, C. D. and Frith, U. Implicit and explicit processes in social cognition. *Neuron*, 60(3):503–510, 2008.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., et al. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, October 2022. OpenAI.

Goodhart, C. A. E. Problems of monetary management: The U.K. experience. In *Papers in Monetary Economics*, volume 1. Reserve Bank of Australia, 1975.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Guerzhoy, M. Barriers to complexity-theoretic proofs that achieving AGI using machine learning is intractable. *arXiv preprint arXiv:2411.06498*, 2024.

Harnad, S. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346, 1990.

Harnad, S. Problems, problems: The frame problem as a symptom of the symbol grounding problem. *Psycoloquy*, 4(34), 1993.

Harnad, S. Symbol ungrounding: What the successes (and failures) of large language models reveal about human cognition. *Frontiers in Robotics and AI*, 11, 2024. PMC11529626.

Hart, B. and Risley, T. R. *Meaningful Differences in the Everyday Experience of Young American Children*. Brookes Publishing, 1995.

Henrich, J. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press, 2015.

Hassabis, D. AI that can match humans at any task will be here in five to 10 years, Google DeepMind CEO says (interview at DeepMind London headquarters). *CNBC*, March 17, 2025. https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html.

Hendrycks, D., Bengio, Y., Marcus, G., Tegmark, M., Schmidt, E., Song, D., et al. A definition of AGI. *arXiv preprint arXiv:2510.18212*, October 2025. 33-author consortium; framework grounded in Cattell–Horn–Carroll cognitive theory.

Heyes, C. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Belknap Press of Harvard University Press, 2018.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

International Energy Agency. Electricity 2025: Analysis and forecast to 2027. *IEA Report*, 2025. https://www.iea.org/reports/electricity-2025.

Kambhampati, S., Valmeekam, K., Guan, L., Stechly, K., Verma, M., Bhambri, S., Saldyt, L., and Murthy, A. LLMs can't plan, but can help planning in LLM-Modulo frameworks. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. https://arxiv.org/abs/2402.01817.

Khoja, A. and Hiscott, L. AGI's last bottlenecks. *AI Frontiers*, October 2025. https://ai-frontiers.org/articles/agis-last-bottlenecks.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, June 2024.

Karbowski, J. Thermodynamic constraints on neural dimensions, firing rates, brain temperature and size. *Journal of Computational Neuroscience*, 27(3):415–436, 2009.

Lawsen, A. Comment on *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. *arXiv preprint arXiv:2506.09250*, June 2025. Open Philanthropy. Version 1 of this preprint additionally listed Claude Opus (Anthropic) as a co-author; v2 lists Lawsen alone.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Lakoff, G. and Johnson, M. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, 1999.

Lewis, M. and Mitchell, M. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*, 2024.

Levy, W. B. and Calvert, V. G. Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proceedings of the National Academy of Sciences*, 118(18):e2008173118, 2021.

MindStudio. ARC-AGI-3 results: GPT-5.4, Claude Opus 4.6, and Gemini 3.1 all score 0%. https://www.mindstudio.ai/blog/arc-agi-3-results-frontier-models-score-zero, April 2026.

Mirzadeh, S. I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, October 2024. Published as ICLR 2025 conference paper.

Marcus, G. Deep learning is hitting a wall. *Nautilus Magazine*, March 2022. https://nautil.us/deep-learning-is-hitting-a-wall-238440.

Marcus, G. How o3 and Grok 4 accidentally vindicated neurosymbolic AI. *Marcus on AI*, 2025.

Mu, N. The myth of data inefficiency in large language models. https://www.normanmu.com/2025/02/14/data-inefficiency-llms.html, 2025.

Farkaš, I., Vavrečka, M., and Wermter, S. Will multimodal large language models ever achieve deep understanding of the world? *Frontiers in Systems Neuroscience*, 19:1683133, November 2025. DOI: 10.3389/fnsys.2025.1683133.

Poetiq. Shattering ARC-AGI-2 state of the art at half the cost. https://poetiq.ai/posts/arcagi_verified/, March 2026.

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., and Dean, J. The carbon footprint of machine learning training will plateau, then shrink. *IEEE Computer*, 55(7):18–28, 2022.

Pavlick, E. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041, 2023.

Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Phan, L., Gatti, A., Han, Z., Li, N., Yue, S., Wang, A., Hendrycks, D., et al. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*, 2025.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., and Roy, D. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668, 2015.

Rumelhart, D. E. and McClelland, J. L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, 1986.

Sagan, C. *Cosmos*. Random House, 1980.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2023.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. The curse of recursion: Training on generated data makes models forget. *Nature*, 631:755–759, 2024.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Symbolica AI. From 0% to 36% on day 1 of ARC-AGI-3. *Symbolica Blog*, March 2026. https://www.symbolica.ai/blog/arc-agi-3.

Spelke, E. S. and Kinzler, K. D. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Maslej, N., Fattorini, L., Perrault, R., et al. The AI index 2025 annual report. Stanford Institute for Human-Centered Artificial Intelligence, 2025.

Strathern, M. 'Improving ratings': audit in the British university system. *European Review*, 5(3):305–321, 1997.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020.

Tomasello, M. *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., and Rich, P. Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 7:616–636, 2024. https://doi.org/10.1007/s42113-024-00217-5.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. Will we run out of data? Limits of LLM scaling based on human-generated data. *International Conference on Machine Learning*, 2024. https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

Vygotsky, L. S. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *International Conference on Learning Representations*, 2019.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Wolpert, D. H. and Macready, W. G. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.