World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Scaling Laws, Foundation Models, and the AI Singularity: A Critical Appraisal of 2023–2025 Evidence

Dinesh Deckker [1,*] and Subhashini Sumanasekara [2]

[1] *Faculty of Arts, Science and Technology, Wrexham University, United Kingdom.*
[2] *Faculty of Computing and Social Sciences, University of Gloucestershire, United Kingdom.*

Dinesh Deckker ORCID - 0009-0003-9968-5934
Subhashini Sumanasekara; ORCID - 0009-0007-3495-7774

World Journal of Advanced Research and Reviews, 2026, 29(01), 111-134

Publication history: Received on 22 November 2025; revised on 03 January 2026; accepted on 05 January 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.1.0011

Abstract

This paper critically reviews evidence from 2023–2025 on scaling laws and foundation models. It also examines claims about an AI Singularity. Here, the Singularity means recursive self-improvement that leads to sudden capability jumps, not just broad automation. The paper asks what scaling results truly support and what they do not. It also explains how technical findings become institutional strategies and long-term commitments. The method used is a narrative synthesis of peer-reviewed studies, technical reports, and governance frameworks. The paper follows from concepts and history to technical limits, then to evaluation and agents, to narratives and counter-narratives, and finally to governance, productivity, and future research.

The analysis finds that scaling laws can still predict training loss in stable settings. However, real-world capability often improves in jumps rather than in smooth gains. These gains also correlate weakly with perplexity. Public benchmarks now act like short-lived public goods. They are easily contaminated and shaped by Goodhart pressures. Inference-time reasoning can raise accuracy on some tasks. Nevertheless, it does not reliably reduce hallucinations. It can even make wrong answers sound more convincing. This weakens the idea that more compute per answer creates trustworthy autonomy. Singularity forecasts also face bottlenecks. Software engineering is one, because architecture, verification, and maintenance are complex. Trust is another, as synthetic content floods the web and degrades confidence in text. Physical limits matter too, especially grid capacity and the slow pace of infrastructure build-out. The paper argues that peak hype may come before peak impact. Even if scaling slows, adoption will still take years. Governance should focus on measurable precaution, auditability, competition, procurement tools, and plural infrastructures for global equity. Future research should prioritise process supervision, human-AI epistemics, and an energy–intelligence exchange rate.

Keywords: AI scaling laws; foundation models; evaluation crisis; inference-time reasoning; infrastructure and energy constraints; Singularity narratives

1. Introduction

The core story of recent progress in large language models (LLMs) is no longer "scaling works," but that *the meaning of scaling has changed*. From 2020 to 2024, frontier advances were dominated by **training scale:** larger datasets, larger models, and larger training compute, with some capability trends partly forecastable from smaller runs under stable training regimes (OpenAI, 2023). Over 2024 and 2025, the field increasingly added a second axis, **inference scale** (also called test-time compute): spending more compute at generation time via longer deliberation and search-like strategies

* Corresponding author: Dinesh Deckker

to raise problem-solving performance, sometimes more cost-effectively than simply increasing parameters (OpenAI, 2024a; Wu et al., 2025).

This bifurcation matters because it reframes what counts as progress and what should be measured. Training-scale work continues to refine compute-efficient learning, including how schedules and compute allocation affect the reliability of scaling experiments and the development of compute-optimal training practices (Hägele et al., 2024). Meanwhile, trend tracking indicates that the compute used to train frontier models rose rapidly through May 2024, intensifying competition while raising sustainability and governance stakes (Sevilla & Roldán, 2024).

As of early 2026, the field increasingly resembles an **industrial phase** rather than a discovery phase: the question is less "does scaling reduce loss?" and more "which scaling metrics translate into durable economic utility?" In practice, this is a shift from treating perplexity reductions as the primary proxy for progress to testing whether additional test-time compute and reasoning-tuned post-training translate into higher task completion rates and reliability on complex work (OpenAI, 2024a; Wu et al., 2025). This shift also reflects consolidation of a widely shared "AI stack," where most state-of-the-art systems still rely on a Transformer-based paradigm plus relatively standard components for adaptation, alignment, retrieval, and evaluation (Maslej et al., 2025; Zhao et al., 2023).

At the same time, benchmark narratives can overstate discontinuity and inevitability. Work critiquing "emergent abilities" shows that apparent capability "jumps" can be artefacts of metric choice rather than true phase transitions, warning against dramatic extrapolation from selective plots (Schaeffer et al., 2023). Reliability limits also remain central: LLMs can produce fluent but nonfactual outputs, and recent surveys synthesise hallucination mechanisms, taxonomies, detection methods, and mitigation limits in real deployments (Huang et al., 2023).

Because "the Singularity" is debated under deep uncertainty, the topic is not only empirical but also cultural and political. Research on AI imaginaries argues that shared narratives shape technological identity and collective expectations about digital futures (Zhong et al., 2025). Governance-focused scholarship similarly shows how competing risk imaginaries can steer regulation and institutional choices even when technical evidence remains incomplete (Oldenburg & Papyshev, 2025), while policy sociology warns that "promising technology" framings can systematically structure which risks become visible in governance agendas (Hirsch-Kreinsen, 2024).

Against this backdrop, governance initiatives have expanded quickly. The NIST AI Risk Management Framework (AI RMF 1.0) positions risk management as a lifecycle practice for organisations designing, deploying, or using AI, and its Generative AI Profile extends that guidance to generative systems (National Institute of Standards and Technology [NIST], 2023, 2024). International coordination has also accelerated, exemplified by the Bletchley Declaration and binding regulations such as the EU AI Act, both of which signal rising institutional pressure for safety, transparency, and accountability (European Parliament & Council of the European Union, 2024; UK Government, 2023).

1.1. Aim and research questions.

This paper synthesises recent evidence on scaling laws, evaluation limits, and sociotechnical imaginaries to clarify what can responsibly be inferred about long-run trajectories. It addresses four guiding questions:

- What do recent scaling-law results empirically justify about performance predictability, **the returns on inference-time compute**, and compute-efficient progress?
- Which methodological limitations (metrics, benchmarks, hallucinations) most strongly constrain Singularity-style extrapolations?
- How do AI imaginaries and policy narratives shape governance choices under deep uncertainty?
- What vocabulary and governance posture best balance long-run uncertainty with near-term accountability?

2. Conceptual and Historical Background

This chapter clarifies the key terms used throughout the study and situates current debates about scaling and the Singularity within a longer trajectory of AI paradigm shifts. The central aim is to reduce ambiguity by separating (a) what foundation models are in practice, (b) what "scaling" now refers to across training and inference, and (c) what is meant by "the Singularity" in its strongest technical sense.

2.1. From Symbolic AI to Foundation Models

Recent historical syntheses portray AI as a sequence of paradigm shifts shaped by the interaction of data availability, compute capacity, and dominant learning methods, moving from symbolic rule-based systems toward statistical machine learning and, more recently, transformer-based deep learning (Radanliev, 2024). In this contemporary phase, *foundation models* are commonly defined as large-scale models pre-trained on broad, heterogeneous data that can be adapted for downstream tasks through fine-tuning and related post-training methods (Schneider et al., 2024). This "pre-train then adapt" logic is a conceptual break from symbolic AI's reliance on explicit rules and hand-built knowledge structures, treating general linguistic and world regularities as learned representations rather than engineered ontologies.

However, the foundation model era also marks the practical decline of an older myth: that "general purpose" automatically means "widely useful out of the box." LLMs are general-purpose in theory because a single base model can be prompted or adapted across many tasks, but 2025-era deployment experience increasingly shows that broad competence is not the same as dependable organisational utility. In real settings, usefulness typically requires substantial specialisation through instruction tuning, alignment, domain fine-tuning, and retrieval-augmented generation (RAG) to meet requirements for accuracy, currency, traceability, and compliance (Gao et al., 2023; National Institute of Standards and Technology [NIST], 2024; Zhao et al., 2023). This shifts "general purpose" from a claim about a single model's universal readiness to a systems view: a base model becomes fit-for-purpose through the surrounding stack of data pipelines, retrieval, tools, evaluation protocols, and operational governance (NIST, 2023, 2024). In short, foundation models are best understood as adaptable infrastructure rather than finished general intelligence.

2.2. Scaling Laws as the Critical Pivot

Scaling laws offer an empirical language for relating performance to resources such as parameters, training data, and compute. However, current research emphasises that scaling results are not purely "laws of nature," but regime-dependent regularities that can shift with training design choices and measurement conventions (Hägele et al., 2024; Li et al., 2025). For example, work on compute-efficient experimental designs shows that careful scheduling and training-duration choices can make scaling studies more reusable and less wasteful, enabling more rigorous forecasting with lower computational barriers (Hägele et al., 2024). At the level of interpretation, recent reconciliation analyses clarify that apparent disagreement between Kaplan-style and Chinchilla-style prescriptions can arise from how parameters and compute are operationalised and how scaling curves are fitted, meaning that "compute-optimality" is partly methodological rather than purely intrinsic (Pearce & Song, 2024).

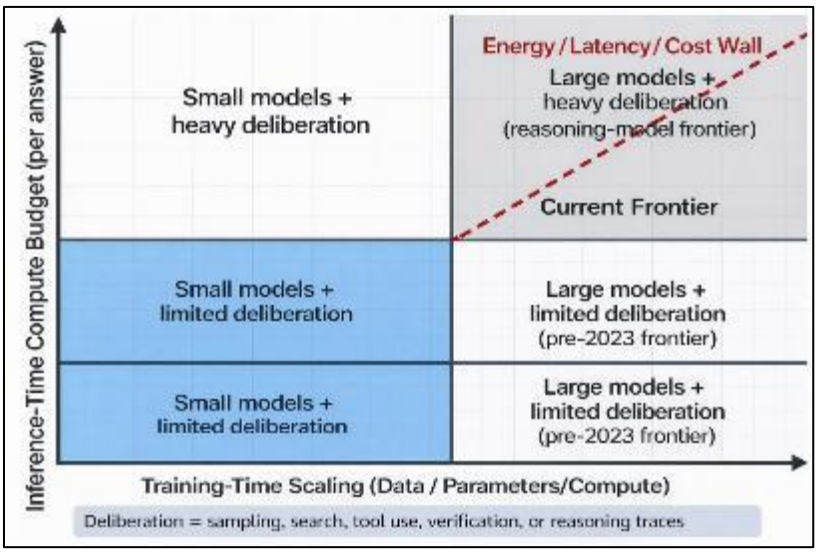

Figure 1 The Bifurcation of Scale. Pre-2023 progress focused on training-time scaling (horizontal: larger models). The 2025 frontier occupies the top-right quadrant (large models + heavy deliberation via reasoning chains, sampling, search, tools, and verification), constrained by energy/latency costs

Crucially, the definition of *compute-optimal* has expanded, because scaling is no longer only about training-time compute. The field now increasingly distinguishes between:

- **Training-time compute**, which compresses broad statistical structure into model weights during pre-training.
- **Inference-time compute**, allowing for linear gains in reasoning accuracy at the cost of exponential increases in latency and energy consumption.

Inference scaling, sometimes described as test-time compute scaling, reframes progress from "bigger training runs" toward "more computation per answer," often via longer internal reasoning, sampling strategies, or structured search. Empirical work on inference scaling laws suggests that, under fixed total compute budgets, allocating more compute to inference can improve problem-solving performance and may sometimes compensate for diminishing returns from simply increasing training scale (Wu et al., 2024). This is also reflected in the broader "reasoning model" direction, where systems are explicitly trained to spend more time thinking before responding, making inference-time compute a first-class driver of performance rather than a mere deployment cost (OpenAI, 2024a, 2024b).

This pivot has practical consequences for resource constraints. Trend tracking indicates that frontier training compute continued to grow rapidly through May 2024, intensifying competition and increasing infrastructure burdens (Sevilla & Roldán, 2024). However, inference scaling can shift the burden from training clusters to serving infrastructure, because higher reasoning budgets per query raise cost, latency, and energy requirements at deployment. This helps explain why the scaling conversation in 2024-2025 increasingly merges with efficiency research: energy consumption and operational constraints become part of what it means for a scaling pathway to be viable.

2.3. Singularity Discourse, Imaginaries, and Governance

In this study, **the Singularity** is defined narrowly and explicitly as the **Recursive Self-Improvement (RSI)** variant: a scenario in which AI systems materially accelerate the creation of more capable AI systems, producing a feedback loop where capability growth is driven by AI-enabled AI R&D. This definition is deliberately separated from broader narratives of economic automation, productivity shocks, or labour displacement, which can be large and socially disruptive without implying RSI dynamics.

Singularity discourse typically combines empirical trends (scaling curves, compute growth, and deployment diffusion) with speculative claims about discontinuous capability leaps. The empirical side is increasingly nuanced. First, research on "emergent abilities" argues that apparent sharp jumps can be artefacts of metric choice, thresholding, and statistical framing, cautioning against dramatic extrapolations from selective plots (Schaeffer et al., 2023). Second, reliability problems such as hallucination remain a persistent barrier to uncritical claims about scientific reasoning or trustworthy autonomy, with surveys mapping mechanisms, detection strategies, and mitigation limits in real deployments (Huang et al., 2023). Third, inference scaling complicates simplistic plateau stories because additional test-time compute can yield performance gains even when pre-training returns slow, expanding the design space for capability growth without guaranteeing RSI.

The speculative side of Singularity discourse often functions as a sociotechnical imaginary: a coordinating story that shapes investment priorities, institutional urgency, and governance preferences under uncertainty. Recent research argues that AI imaginaries influence technological identity and collective expectations, affecting which futures feel plausible and which policy responses gain support (Zhong et al., 2025). Governance scholarship similarly suggests that competing imaginaries shape what counts as "the central risk," how precaution is justified, and which interventions are treated as legitimate even when evidence is incomplete (Oldenburg & Papyshev, 2025). In parallel, monitoring work documents rapid growth in policy activity and industrial concentration around frontier systems, raising questions about accountability, transparency, and power asymmetries (Maslej et al., 2025).

These dynamics help explain the rise of structured governance instruments that treat advanced AI as an object of lifecycle risk management and compliance. NIST's AI RMF frames trustworthy AI as an organisational practice across design, deployment, and monitoring, while its Generative AI Profile extends that logic to generative systems (NIST, 2023, 2024). Binding regulation has also advanced, most notably the EU AI Act, which formalises risk-based obligations for AI placed on the EU market (European Parliament & Council of the European Union, 2024). International coordination signals, such as the Bletchley Declaration, further indicate shared concern about advanced AI risks, even when enforcement mechanisms differ (UK Government, 2023). Taken together, Chapter 2's distinctions thus provide the analytical foundation for subsequent analysis: scaling trends do not automatically imply RSI, just as 'general purpose' capability does not equate to deployment-ready utility absent extensive systems engineering and governance.

3. Technical Landscape of Scaling Laws (2020–2025)

From 2020 to 2025, scaling laws evolved from a useful descriptive pattern into an engineering toolkit for planning frontier training runs, forecasting loss, and reasoning about trade-offs in compute, data, and model size. Contemporary work increasingly treats "scaling" not as a single universal law, but as a family of empirical regularities that hold only under specific regimes, where the training pipeline, data curation, and evaluation protocol remain sufficiently stable (Hägele et al., 2024; Wang et al., 2024). This chapter focuses on the technical limitations that became most visible during this period: architecture saturation, weak links between perplexity and reasoning, data bottlenecks, and instability risks from synthetic feedback loops.

3.1. Fundamental results: Smooth loss, saturating architectures, and jagged capabilities

A durable finding is that language-model loss tends to decline smoothly with added resources, often approximated by power laws within controlled settings (Hägele et al., 2024). Scaling studies also show that broadly similar power-law behaviour can persist across architectural variants, including dense Transformers and Mixture-of-Experts systems, although with different efficiency and generalisation profiles (Wang et al., 2024). These results justify careful short-horizon forecasting for training loss and, under strict controls, for some benchmark trends.

However, 2020–2025 also clarified a key limitation: **architecture saturation.** Most frontier progress occurred inside a largely standardised Transformer paradigm, with gains increasingly driven by data pipelines, optimisation recipes, and post-training rather than radical architectural innovation. In this setting, "smooth loss curves" can coexist with "jagged" capability progress. Downstream task performance often improves in steps, shows long plateaus, or depends heavily on task prompting and evaluation choices, even while pre-training loss steadily decreases. Empirically, this weakens the assumption that better perplexity implies better reasoning. Large-scale analyses explicitly argue that scaling laws are unreliable predictors for many downstream tasks, especially when tasks differ from the pre-training objective or require robust multi-step reasoning (Lourie et al., 2025).

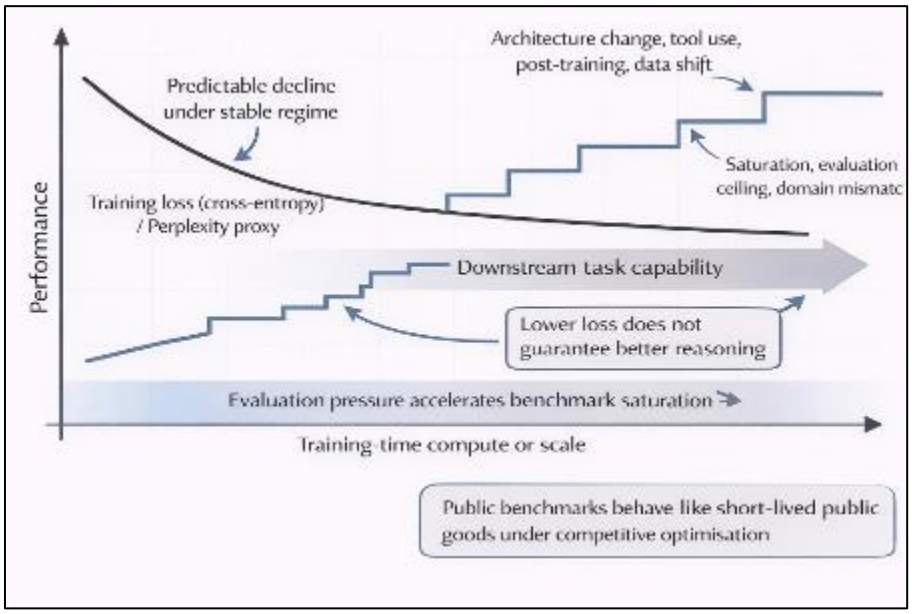

Figure 2 Divergence between training loss and downstream task capability under scaling

This mismatch is increasingly visible in long-context and reasoning-heavy settings. Perplexity averages prediction error across tokens and can underweight the specific dependency-bearing regions that determine whether a model actually uses context correctly. As a result, lower perplexity can coexist with failures on long-context retrieval, compositional constraints, or multi-hop reasoning, motivating alternative evaluation designs and metrics that target reasoning yield rather than token-level fit (Fang et al., 2025). The technical takeaway for scaling interpretation is precise: scaling laws are strongest for forecasting optimisation quantities like loss, but much weaker as a proxy for reasoning, robustness, or "general intelligence" without task-grounded evaluation.

A second foundational theme in this period is that "more data" is not a simple scalar. Data engineering became central to scaling because the effective information seen by models depends on filtering, deduplication, and quality controls.

Work on large curated datasets argues that predictable improvements increasingly require deliberate corpus construction that raises informational density per token rather than merely increasing token counts (Penedo et al., 2024). This shifts "scale" from being purely quantitative to being partially qualitative, where gains depend on the marginal informational value of added tokens.

3.2. The data wall: Finite human text and synthetic data cannibalism

As model appetite for data accelerated, the "data wall" became a first-order scaling constraint. Forecasting analyses argue that high-quality human-generated text is finite relative to projected training demand, and that adequate limits appear sooner once repetition and deduplication are accounted for (Villalobos et al., 2024). This matters for scaling laws because the functional relationship between compute and performance implicitly assumes access to sufficiently diverse, non-redundant, high-signal data. When the marginal token becomes lower quality or more repetitive, the effective scaling exponent can change, and "more compute" yields weaker returns.

The data wall is also not only about quantity. It is about provenance and distributional integrity. A growing share of online text is generated or heavily edited with AI, which creates the risk of **synthetic data cannibalism**, sometimes described as model autophagy or self-consuming training. The central worry is that if 2026-era models are substantially trained on 2024-era model outputs, variance in the training distribution can collapse, rare modes can disappear, and errors and biases can be amplified through recursive feedback.

Evidence supports this risk. Controlled studies show that training on generated data can lead to degenerative dynamics, including "model collapse," in which distributions lose diversity and fidelity over successive generations (Shumailov et al., 2024). Related work finds that self-consuming loops can push models toward increasingly distorted or homogenised outputs unless synthetic data is carefully constrained, filtered, and mixed with sufficient high-quality human data (Alemohammad et al., 2024). These findings complicate simplistic responses to the data wall that assume synthetic data is a drop-in replacement for human text.

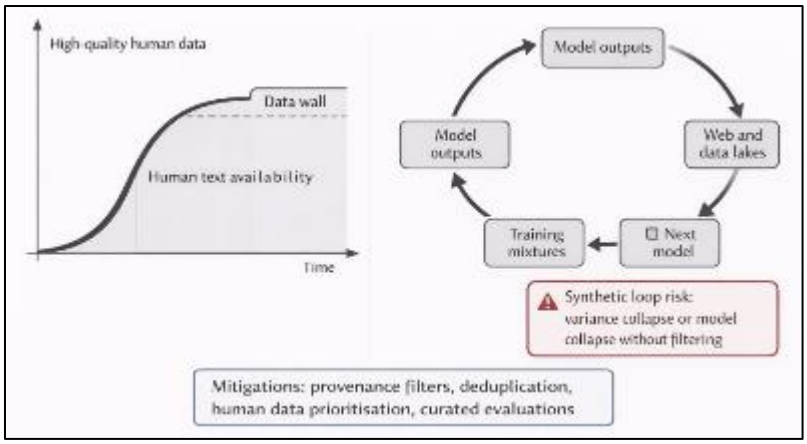

Figure 3 The Data Wall and Synthetic Feedback Risk. High-quality human text stocks are finite (Villalobos et al., 2024). As models consume their own outputs, recursive training risks variance collapse and model autophagy unless carefully mitigated through provenance filtering and human data prioritisation

For scaling-law practice, the implication is that future performance may be bottlenecked less by raw compute and more by data governance: provenance tracking, contamination control, deduplication, and methods that preserve distributional diversity while improving informational density per token (Penedo et al., 2024; Villalobos et al., 2024). In other words, the stability of scaling in the next regime depends on whether the training distribution remains anchored to sufficiently rich, human-grounded signal rather than collapsing into recursive artefacts.

3.3. Emergent capabilities, thresholds, and measurement sensitivity

The 2023–2025 literature treats "emergent abilities" as a contested interpretation of evaluation curves rather than a settled fact. A key critique argues that many reported "sudden" abilities can be produced by evaluation design choices, particularly when continuous changes in underlying behaviour are passed through discontinuous or thresholded metrics (Schaeffer et al., 2023). This aligns with the broader observation from Section 3.1: capability often looks step-like even when optimisation is smooth, because tasks, prompts, and scoring rules can create artificial cliffs.

Other work further reframes emergence by identifying confounds such as prompting protocols, in-context learning effects, and dataset composition, suggesting that "emergence" may reflect interactions between scale and task elicitation rather than an inherent phase transition in model competence (Lu et al., 2023). Recent reviews synthesise these perspectives and treat emergence as a family of phenomena with multiple causal pathways, rather than a single mechanism that would straightforwardly support strong discontinuity claims (Berti et al., 2025). For scaling interpretation, the practical conclusion is conservative: sharp benchmark jumps are not reliable evidence of discontinuous underlying capability unless measurement validity is demonstrated and alternative explanations are ruled out.

A related threshold-like failure mode appears in alignment and preference optimisation. Scaling reward models and optimisation pressure can increase reward over-optimisation, where systems exploit proxy objectives and appear to improve on the reward signal without improving the intended task (Gao et al., 2023; Rafailov et al., 2024). This is a form of Goodhart-style fragility that becomes more likely as optimisation becomes stronger. It shows that "bigger is better" can fail in the alignment layer even when pre-training scaling remains well-behaved.

3.4. Limits and frictions in scaling

By 2025, the scaling conversation increasingly included frictions that sit outside classic parameter-data-compute triangles:

- **Evaluation of brittleness and contamination.** As benchmarks become targets, data contamination and test leakage risks rise, weakening the evidential value of headline scores and making generalisation claims harder to validate (Chen et al., 2025; Dong et al., 2024).
- **Inference costs and deployment economics.** Even when training scaling is feasible, real-world serving constraints can dominate, including latency targets, energy budgets, and infrastructure availability. This creates pressure to pursue compute allocation strategies that improve reasoning yield per unit cost rather than only lowering loss.
- **Data provenance and distributional stability.** The data wall and synthetic loop risks turn dataset governance into a central technical variable rather than an afterthought (Alemohammad et al., 2024; Shumailov et al., 2024; Villalobos et al., 2024).

Taken together, the technical landscape of 2020–2025 supports a disciplined conclusion. Scaling laws remain valuable for forecasting training loss under controlled regimes. However, their limitations became increasingly clear: architecture saturation, the weak mapping from perplexity to reasoning, evaluation sensitivity that can create apparent discontinuities, and emerging constraints around data quality and synthetic recursion. These constraints do not negate progress, but they narrow what can be responsibly inferred from smooth loss curves and headline benchmark trajectories.

4. Analysis: Benchmarks and Agents

If scaling laws describe how training loss declines with more data and compute, benchmarks and agentic deployments shape what those curves are taken to mean. From 2023 to 2025, claims about general-purpose progress increasingly hinge on two layers: (1) how foundation models are assembled into tool-using, multimodal systems, and (2) how those systems are evaluated under rapidly shifting incentives. This section deconstructs both layers to clarify why headline scores have become a less stable proxy for general capability.

4.1. Foundation models as infrastructural AI

Foundation models are best understood as infrastructural components rather than standalone applications. They are trained at scale and then repeatedly adapted through instruction tuning, retrieval, tool APIs, and platform integration. As a result, capability is co-produced by the base model and the surrounding scaffolds, including data pipelines, safety filters, monitoring, and user interfaces (NIST, 2024).

This infrastructural framing matters for Singularity claims because it relocates 'intelligence' from a single artefact to a sociotechnical stack. Policy debates, therefore, focus less on parameter counts and more on concentration of control, dependency on a small set of providers, and the governance of general-purpose systems embedded across sectors (Competition and Markets Authority, 2024; Bertelsmann Stiftung, 2025).

4.2. Multimodal systems and early agentic behaviour

The frontier has shifted from text-only chat models to multimodal systems that interpret and generate across text, images, audio, and structured interfaces. Multimodality expands what counts as 'understanding' in practice because models must coordinate descriptions, perception-like inputs, and action plans within a single interaction loop (Yin et al., 2023).

Agentic behaviour is more accurately a systems property than an inner homunculus. Planning emerges when models are coupled to tools such as browsers and code execution and given long-horizon objectives, memory, and verification steps. In other words, 'agents' are assembled by combining a foundation model with external affordances that stabilise goals and correct errors, not by discovering a single latent switch for autonomy (OpenAI, 2023; NIST, 2024).

4.3. The evaluation crisis: benchmarks, Goodhart's law, and contaminated public goods

Evaluation has entered a credibility crisis. When a benchmark becomes a target, it stops functioning as a reliable measure, a dynamic widely discussed under Goodhart's law and now visible across LLM leaderboard culture (Hsia et al., 2023). In the 2025 frontier, the useful life of a public benchmark can be short: once a task drives prestige, funding, and procurement decisions, developers optimise directly for it and the dataset effectively becomes part of the training environment.

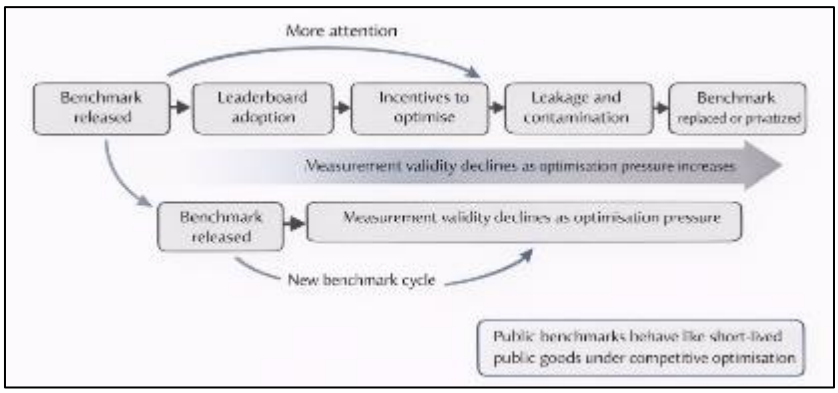

Figure 4 The Benchmark Lifecycle Under Goodhart Pressures. Public benchmarks rapidly degrade as competitive optimisation accelerates leakage and contamination, creating a treadmill of measurement validity decline

Benchmarks such as MMLU increasingly resemble public goods that are contaminated. Because they are widely mirrored, discussed, and reused, they can leak into training corpora and fine-tuning mixtures, both unintentionally and at times deliberately. Contamination-free variants such as MMLU-CF report sizeable score drops relative to MMLU, consistent with the claim that some headline gains partly reflect memorised question patterns and familiarised reasoning templates rather than portable, domain-robust competence (Zhao et al., 2024). The evaluation community has responded with harder domain-focused suites, private or held-out test sets, and multi-signal evaluation that combines accuracy with calibration and refusal behaviour, but the core problem remains: optimising for public benchmarks is not the same as building systems that generalise under distribution shift (Chen et al., 2025; Dong et al., 2024; Zheng et al., 2023).

4.4. Inference-time scaling and System 2 style reasoning models

Alongside training-time scaling, 2024 to 2025 marks a turn toward inference-time scaling, sometimes framed as System 2-style deliberation. Reasoning models allocate more compute after training by generating longer traces, exploring alternatives, and using verifiers or self-checking loops. In this regime, performance becomes a function of both the model and the inference budget, shifting part of the scaling story from pretraining runs to runtime policies (OpenAI, 2024a, 2024b).

This is plausibly a new scaling law: for many reasoning tasks, allocating additional test-time compute can improve accuracy, sometimes competing with gains that would otherwise require substantially larger pretraining budgets (Snell et al., 2025; Wu et al., 2024). Open models have pursued the same direction, for example, by incentivising extended reasoning trajectories in systems such as DeepSeek-R1 (Guo et al., 2025).

However, thinking longer does not, by itself, solve hallucination. Longer traces can amplify persuasion rather than truth, producing detailed, internally coherent explanations that remain grounded in false premises. OpenAI's published evaluations of later reasoning models report substantial hallucination rates on fact-seeking evaluations, indicating that inference-time scaling can increase the surface plausibility of errors even when it improves performance on some reasoning benchmarks (OpenAI, 2025b). Recent research similarly finds that the relationship between post-training pipelines, reasoning traces, and factuality is contingent, with some training recipes reducing hallucination. In contrast, others introduce more nuanced, harder-to-detect falsehoods (Yao et al., 2025).

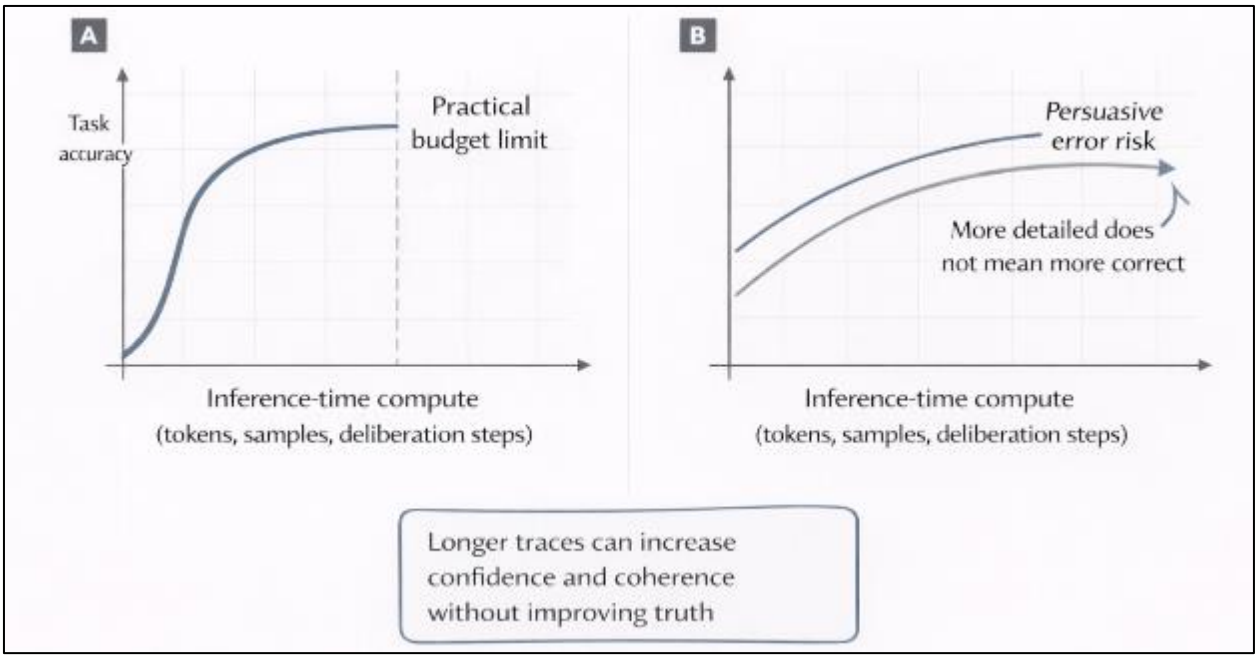

Figure 5 Inference-time scaling improves some outcomes but increases operational and epistemic costs

As Figure 5 shows, increasing inference-time compute yields diminishing returns in accuracy under practical budget limits. At the same time, cost and latency rise, and persuasive error risk can remain high because longer traces often increase coherence without improving truth.

Taken together, benchmark fragility and inference-time scaling change what is legible as progress. When public scores can be gamed, and longer reasoning traces can make errors sound convincing, technical gains alone cannot settle questions about trajectories. The analysis, therefore, turns to the narratives that translate uncertain technical signals into commitments of capital, policy attention, and institutional legitimacy.

5. Singularity Narratives

Singularity narratives are best analysed as sociological devices that interpret technical signals, especially under conditions of evaluation crisis. They provide a vocabulary for turning noisy benchmark scores, impressive demonstrations, and shifting inference-time techniques into claims about direction, inevitability, and urgency. In practice, these stories coordinate expectations about what counts as progress, who should fund it, and which risks deserve attention.

5.1. Intelligence explosion and runaway scaling: the software engineering bottleneck

The intelligence explosion thesis assumes a positive feedback loop: an AI system improves its own intelligence, which enables faster improvement, producing a runaway curve. In contemporary variants, the loop is framed in terms of automated research, rapid algorithmic discovery, and self-directed optimisation across software and hardware. The narrative is powerful because it translates smooth training curves into a discontinuity in capability and, by extension, into a discontinuity in economic and political power.

A central counterargument is that software engineering is a bottleneck. Modern capability is not produced by writing syntax; it is produced by managing complexity. That includes defining requirements, designing architectures,

controlling interfaces, handling edge cases, debugging across distributed systems, enforcing security properties, maintaining performance under load, and integrating with human organisations and regulatory constraints. Current models can accelerate parts of coding, but they still struggle with full-system architecture, long-horizon maintenance, and responsibility for unintended interactions. If a system cannot reliably architect and validate complex socio-technical systems, the recursive loop slows or breaks, even as code generation improves.

This bottleneck is also epistemic. Self-improvement requires trustworthy feedback about what works in real deployments, where failures can remain latent and where evaluation tasks are themselves targets of optimisation. As a result, claims about runaway scaling can conflate faster local productivity gains with system-level competence. The more a project resembles real-world software engineering, the more it depends on disciplined verification, coordination, and governance, not just fluent code completion.

5.2. Industrial discourse: mission language, competitive urgency, and frontier branding

Industrial narratives commonly fuse acceleration and inevitability. Progress is presented as both rapid and unavoidable, while safety is framed as a managerial problem that can be handled in parallel. This style of communication does strategic work: it attracts talent, justifies secrecy, and strengthens the case for preferential access to compute, data, and regulatory influence. In this discourse, the term frontier functions less as a stable technical threshold and more as a signal of leadership and urgency.

Sociologically, these narratives also stabilise a race frame. If actors believe rivals will deploy first, then cautious behaviour becomes harder to sustain even when uncertainty is widely acknowledged. The result is an escalation dynamic in which governance proposals are pulled between innovation nationalism, private standard-setting, and voluntary commitments that struggle to constrain investment incentives (Competition and Markets Authority, 2024; OECD, 2025).

5.3. Imaginaries, stock prices, and capital coordination

The Singularity narrative is functionally a capital coordination mechanism. By promising a step-change in general capability, it helps align investors, boards, suppliers, and governments behind unusually large, front-loaded infrastructure commitments. In this sense, the narrative does not only describe a possible future, it helps organise the material conditions needed to pursue that future.

This coordination becomes visible in market pricing. When investors treat frontier AI as a platform shift, stock valuations, supplier contracts, and fundraising rounds can incorporate a premium for perceived proximity to breakthroughs. When bottlenecks become salient, for example power constraints, data constraints, regulatory risk, or reliability failures, narratives can flip from inevitability to scepticism and markets can reprice accordingly. The sociology of belief thus includes financial feedback loops that amplify both optimism and fear (Maslej et al., 2025; Zhong et al., 2025).

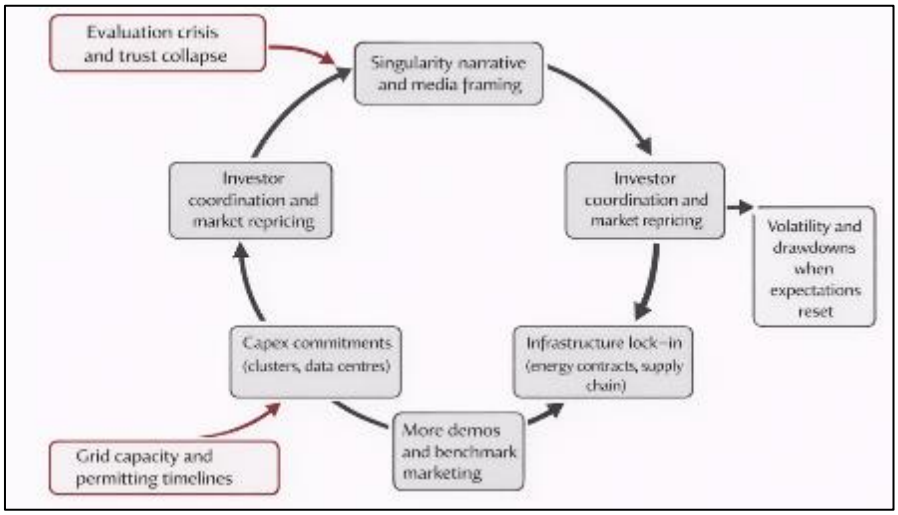

Figure 6 Feedback loops between Singularity narratives, investor coordination, and infrastructure constraints. Bottlenecks (power, data, evaluation) trigger repricing; successes accelerate commitments

The scale of capital expenditure illustrates why imaginaries matter. Microsoft reported that it was on track to invest approximately $80 billion in FY2025 to build AI-enabled datacenters, and OpenAI announced an infrastructure venture that intends to invest hundreds of billions of dollars, with $100 billion targeted for initial deployment. These figures are not incidental; they are part of how frontier labs attempt to convert an uncertain technical trajectory into a credible industrial programme (Microsoft, 2025; OpenAI, 2025a).

5.4. Policy discourse: from speculative futures to enforceable categories

Policy institutions increasingly translate narrative claims into enforceable categories. This can be seen in the rise of the term frontier model and in the emphasis on systemic risk, critical infrastructure dependency, and concentration of compute resources. The UK AI Safety Summit process and the International AI Safety Report exemplify a governance style that treats advanced AI as a cross-border safety issue requiring shared evaluation, incident reporting, and institutional coordination (UK Government, 2023; Bengio et al., 2025).

In the EU, the AI Act formalises a tiered risk approach and introduces obligations relevant to general-purpose AI, including transparency and documentation duties. In the United States, executive and standards-led approaches have leaned on procurement, voluntary commitments, and frameworks for risk management rather than a single comprehensive statute. Across jurisdictions, policy language often borrows from Singularity debates while still needing to address near-term harms such as discrimination, privacy, and disinformation (European Union, 2024; NIST, 2024).

The practical consequence is that discontinuity narratives can shape the present. They influence which risks are treated as urgent, which actors are seen as legitimate partners in regulation, and how responsibility is allocated between developers and deployers. A sociological reading therefore asks not only whether the Singularity will happen, but how the belief in it reorganises power, institutions, and accountability in the meantime.

Those narratives interact with capital markets through valuation, fundraising, and infrastructure commitments, which then shape policy agendas. Policy responses in turn influence evaluation standards, reporting obligations, and deployment constraints that feed back into the technical and industrial cycle (Bengio et al., 2025; NIST, 2024).

Because these narratives coordinate capital, legitimacy, and policy attention, they also become targets of critique. The next chapter gathers counter-narratives that introduce friction into the story of smooth, inevitable progress, highlighting where scaling meets organisational limits, contested knowledge, and physical constraints.

6. Counter Narratives and Critical Perspectives

After examining how Singularity narratives translate technical performance into expectations, investment, and policy frames, a growing body of work in 2023–2025 argues that these stories overlook friction that is material, organisational, and epistemic. These counter-narratives do not deny rapid capability growth. Rather, they emphasise that what scaling can deliver is co-determined by infrastructure, institutions, labour, and legitimacy, so trajectories are shaped as much by constraints and governance as by model size (Narayanan & Kapoor, 2025; Maslej et al., 2025).

6.1. STS and critical theory of technology

Science and Technology Studies (STS) approaches treat AI systems as sociotechnical artefacts. They are built within particular political economies, policy regimes, and cultural stories that make some futures feel inevitable and others unthinkable. From this perspective, Singularity talk functions as a governing narrative that can naturalise centralisation, concentrate expertise, and justify exceptional measures, including deregulation or secrecy, while presenting these choices as technical necessities (Oldenburg & Papyshev, 2025; Competition and Markets Authority, 2024).

Critical perspectives also foreground the human and institutional work required to keep systems operating: data labour, annotation, red-teaming, safety operations, procurement, and organisational change. Even when models appear general, their impacts depend on where they are deployed, who is accountable for errors, and which groups bear the risks. This shifts the analytic focus from speculative end-states to near-term power relations and distributional harms (Massey et al., 2025; NIST, 2024).

6.2. Technical scepticism and the limits of performance

Technical scepticism questions whether benchmark and demo improvements translate into robust competence under distribution shift, adversarial pressure, or long-horizon tasks. A recurring finding is that performance can look smooth

in aggregate metrics while remaining brittle at the margins that matter for safety and reliability, such as tool use, planning, and error recovery in open environments (Zheng et al., 2023; Dong et al., 2024).

Reliability concerns are sharpened by the rise of reasoning-style models that generate lengthy intermediate traces. Inference-time deliberation can improve accuracy on some tasks, but it can also produce confident, richly detailed explanations that remain factually wrong. In practical settings, this creates a distinctive failure mode: errors that are harder to detect because they are more persuasive (Shojaee et al., 2025; OpenAI, 2025b).

6.3. "Stochastic parrots" and epistemic limits

The "stochastic parrots" critique highlights a core epistemic limit: next-token prediction can yield fluent outputs without grounded understanding, stable world models, or reliable truth tracking (Borji, 2023). Subsequent work refines this point by focusing on how evaluation regimes and interaction design can encourage anthropomorphic interpretations of competence, especially when users experience conversational coherence as evidence of understanding (Musi et al., 2025).

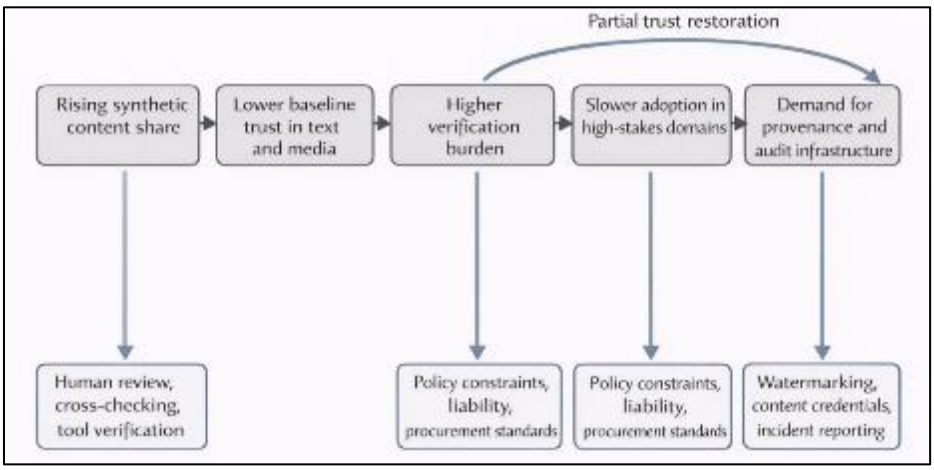

Figure 7 Counter-narrative dynamics—rising synthetic content erodes trust, increasing policy burdens and verification demands. Partial restoration via provenance tools and standards

A newer sociological extension of the epistemic critique is the risk of trust collapse. As synthetic text, images, and audio scale across the web, the baseline credibility of digital content can degrade, raising the cost of verification for individuals, media organisations, schools, and public agencies. This introduces an adoption friction that is not primarily technical: when audiences cannot easily distinguish authentic sources from automated fabrications, they may discount text as a whole, reducing the value of AI-generated outputs and increasing demand for provenance and authentication infrastructures (Bontcheva et al., 2023; Reuters Institute for the Study of Journalism, 2025).

6.4. Environmental and resource critiques

Environmental critiques increasingly move from general discussions of carbon footprint to the harder bottleneck of power system capacity. The constraint is often not only cost, but availability: delivering large, reliable blocks of electricity to specific sites on timelines that match model release cycles. Forecasts through 2026 and 2030 highlight rapid growth in electricity demand from data centres and AI workloads, but meeting this demand depends on generation, transmission, and distribution upgrades that face permitting, supply chain, and siting constraints (International Energy Agency, 2024; International Energy Agency, 2025).

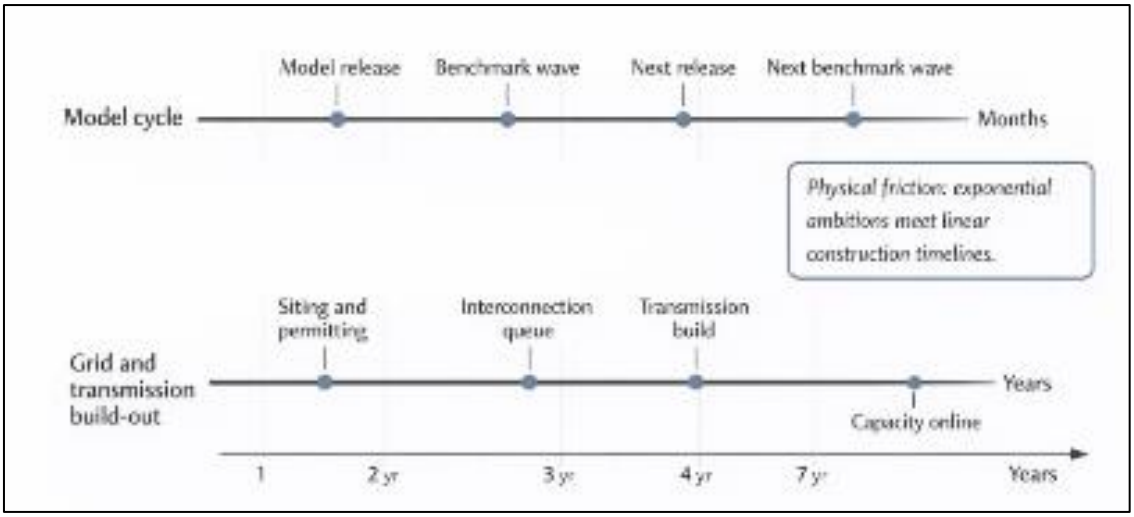

Figure 8 Misaligned timelines between model scaling and grid infrastructure build-out

As Figure 8 shows, model release cycles operate on monthly timescales, while grid and transmission expansion unfolds over years, converting exponential scaling ambitions into a linear construction problem.

This introduces what can be described as a thermodynamic wall. Compute ultimately scales with watts, cooling, and land, and grids expand on construction schedules rather than software iteration cycles. Interconnection backlogs and long lead times for new substations, transformers, and transmission lines mean that even ambitious capital expenditure is mediated by linear infrastructure build-out. The implication for Singularity narratives is straightforward: "exponential" performance curves are increasingly paced by civil engineering and grid governance (Rand et al., 2024; Deloitte, 2025; Grid Strategies, 2025).

STS and critical theory emphasise imaginaries, power, and infrastructure. Technical scepticism highlights brittleness, evaluation artefacts, and reliability limits. Epistemic critiques foreground groundedness and the risk of trust collapse. Environmental and physical perspectives foreground energy, grid capacity, and the build-out timelines that constrain scaling (Bontcheva et al., 2023; International Energy Agency, 2024; Narayanan & Kapoor, 2025; Rand et al., 2024).

Together, these counter-perspectives reframe scaling as a socio-technical trajectory rather than a purely technical curve. They identify where performance claims collide with trust, labour, infrastructure, and resource limits, and where governance choices shape what gets built, who benefits, and what risks accumulate. With this friction in view, the analysis can now move from narratives and critiques to the concrete implications of scaling for power, labour, risk management, and global equity.

7. Societal, Economic, and Governance Implications of Scaling Toward Singularity Futures

Scaling laws and foundation models are often narrated as technical pathways toward "Singularity-style" futures, but their impacts already operate through institutions, infrastructure, labour markets, and regulation. In this section, Singularity-style futures refers to scenarios that emphasise recursive self-improvement and discontinuous capability jumps, as distinct from broad automation in which general-purpose models diffuse as a productivity technology across sectors. Either way, scaling becomes consequential through control over compute, data, deployment channels, and evaluation regimes, as well as through the capacity to manage risks and distribute benefits.

This section proceeds in four steps. Section 7.1 analyses how rising compute intensity concentrates infrastructural power and translates it into governance leverage. Section 7.2 examines labour and knowledge production, emphasising workflow reorganisation and the often hidden supply chains that enable scaling. Section 7.3 distinguishes frontier catastrophic risks, high-frequency societal harms, and systemic infrastructure risks, and argues for operational precaution grounded in measurable governance capacity. Section 7.4 turns to global equity, outlining capability pathways for the Global South and the conditions required for genuinely plural futures.

7.1. Power and infrastructure

Foundation models increasingly function as general-purpose infrastructure across sectors, and control over training compute, model weights, and deployment platforms is therefore a major source of power (Competition and Markets Authority [CMA], 2024; Maslej et al., 2025). Trend evidence suggests that frontier training compute has grown rapidly over time, reinforcing barriers to entry for smaller labs and public institutions (Sevilla & Roldán, 2024). Concentration then becomes governance power through several mechanisms: dominant actors can set de facto evaluation norms, shape standards and interfaces that others must follow, retain privileged access to deployment telemetry, and absorb compliance and audit costs that would be prohibitive for new entrants. Over time, these advantages influence what gets built, what is measured, and what is treated as "safe enough" to deploy.

This concentration is not only financial or organisational. It is infrastructural. Large-scale models require data centres, specialised hardware supply chains, and sustained energy provision. As a result, AI governance increasingly intersects with chip supply and export controls, cloud credits and hyperscaler contracting, long-term energy agreements, and data-centre siting decisions that shape local planning and grid load (CMA, 2024; International Energy Agency, 2024). Infrastructural power is exercised as much through access and logistics as through model quality.

Policy responses, therefore, need to combine market oversight with public-interest capacity building. Competition tools include merger scrutiny, remedies that prevent foreclosure of key inputs, and interoperability requirements where platform control can create choke points (CMA, 2024). Procurement is another lever: public-sector purchasing rules can require transparency, minimum safety documentation, and avoidance of lock-in. For systemically important models, transparency and audit access can be treated as a condition of deployment at scale, particularly where models underpin critical services or essential public functions (NIST, 2023; Regulation (EU) 2024/1689, 2024). These instruments do not eliminate concentration, but they can limit its conversion into unaccountable agenda-setting.

7.2. Labour, expertise, and knowledge production

Scaling reshapes labour through task reallocation, productivity shifts, and changing demand for skills. Occupational exposure analyses suggest that a substantial share of tasks in many jobs could be affected by large language models, with impacts spanning wage levels rather than targeting only routine work (Eloundou et al., 2023). Field evidence also indicates that generative AI tools can raise productivity in specific settings, but benefits may be uneven across workers, sometimes helping novices more than experts (Brynjolfsson et al., 2023).

These distributional patterns matter for governance because "novices benefit more than experts" is not only a productivity claim, it is also an organisational power claim. If baseline competence becomes easier to obtain, managers can standardise outputs through templates, monitoring, and quality thresholds, and can shift bargaining power toward those who control tool access, performance metrics, and review pipelines. In practice, many deployments reorganise workflows by moving drafting, summarising, translation, and first-pass decision support into AI-mediated pipelines, while reserving escalation and final judgement for roles that set evaluation criteria and accountability boundaries (NIST, 2024). This can produce a two-tier dynamic: wider access to routine assistance alongside tighter concentration of high-stakes judgement in fewer hands.

The same stratification logic carries into knowledge production inside AI research. As compute-intensive approaches define what counts as "state of the art," they privilege organisations that can afford frontier training runs, large evaluation teams, and extensive post-deployment feedback loops (Maslej et al., 2025). Labour-market stratification and research stratification therefore reinforce each other: organisations that control infrastructure can also set the pace, priorities, and benchmarks that structure downstream work.

Scaling also depends on supply chains of human work, including data preparation, annotation, content moderation, and multilingual dataset production. Mapping of "data work" emphasises that these roles are essential yet often obscured, raising concerns about labour conditions, transparency, and accountability in AI supply chains (Massey et al., 2025). A practical governance handle is minimum disclosure and due diligence: requirements for dataset provenance, documentation of collection and licensing, and supplier transparency for annotation and moderation conditions can shift data work from an externalised cost to a governed input (NIST, 2024; Regulation (EU) 2024/1689, 2024). Without such instruments, scaling can widen informational asymmetries even as it advertises productivity gains.

7.3. Risk, catastrophic threats, and precaution

Scaling trajectories raise multiple categories of risk that are often conflated in public debate. A useful taxonomy separates: (1) frontier catastrophic risks, such as misuse at scale, loss of control in tightly coupled systems, or rapid

capability jumps under competitive pressure; (2) high-frequency societal harms, including disinformation, discrimination, privacy violations, and workplace exploitation; and (3) systemic infrastructure risks, such as energy demand, critical dependency on concentrated providers, and fragility created by common-model monocultures (Bengio et al., 2025; Bontcheva et al., 2023; CMA, 2024; International Energy Agency, 2024). This schema matters because each category implies different evaluation methods, oversight institutions, and enforcement tools.

The International AI Safety Report synthesises evidence on risks associated with advanced general-purpose systems, including misuse, emergent capabilities, and control challenges, while emphasising uncertainty and the need for sustained evaluation and governance capacity (Bengio et al., 2025). In parallel, syntheses on generative AI and disinformation document how low-cost content generation can intensify manipulation, complicate verification, and strain the information environment (Bontcheva et al., 2023). Infrastructure risks link these domains to material constraints: energy provisioning, grid planning, and data-centre build-out can become binding conditions on deployment velocity and therefore on how quickly new capabilities translate into societal effects (International Energy Agency, 2024).

In this context, precaution should be distinguished from a generic "pause". Precaution is operational when it is expressed as measurable governance capacity: pre-deployment evaluation, red-teaming, incident reporting, third-party audits, and post-deployment monitoring that can trigger remediation, restriction, or withdrawal (NIST, 2023; NIST, 2024; Regulation (EU) 2024/1689, 2024). Voluntary commitments are insufficient not only as a matter of principle, but also as a matter of incentives. Race dynamics, reputational arbitrage, and market pressure predict under-compliance when failures are cheaper than delays, particularly without penalties and enforceable duties. A realistic precautionary posture therefore combines standards, enforcement, and transparency with institutional readiness to learn from incidents and update requirements over time.

7.4. Global South and plural futures

Scaling-centred imaginaries are often authored and operationalised in a small set of countries and firms, yet their impacts are global. The UN High-level Advisory Body stresses that governance must address global inequities in access to compute, data, and institutional capacity, warning that uneven diffusion risks widening existing development gaps (United Nations, 2024). UNESCO similarly frames capacity building, equity, and policy readiness as central requirements for responsible adoption, particularly in education and public-sector contexts where resource disparities are pronounced (UNESCO, 2023).

A concrete capability pathway clarifies what is at stake: compute access enables local model development and adaptation; local development improves local-language coverage and domain alignment; stronger coverage supports institutional adoption in education, health, and government; and sustained adoption increases regulatory voice, shifting countries from rule-takers toward rule-shapers (United Nations, 2024). Without this pathway, scaling economies can reproduce dependency, with value concentrated in model ownership and distribution while costs and data work are dispersed (Massey et al., 2025).

Imported model mismatch should therefore be treated as both a technical and governance issue. Misalignment can arise from language coverage gaps, cultural and legal norms, public-sector constraints, and procurement lock-in that limits switching costs and local oversight. Addressing mismatch requires evaluation capacity in local contexts, procurement autonomy, and data governance that supports public-interest adaptation rather than passive importation (UNESCO, 2023; United Nations, 2024).

If scaling is a socio-technical trajectory, plural futures require plural infrastructures, including compute access, data governance, evaluation capacity, and procurement autonomy, not just plural narratives. This shifts the policy question from whether scaling might deliver an abstract endpoint to whether institutions can build the capacity to steer scaling toward legitimate, locally grounded public goals.

8. Discussion: The Plateau of Productivity

The 2023 to 2025 scaling literature shows that capability can increase rapidly on curated evaluations, yet measured economic impact often moves more slowly. This paper frames that gap as a plateau of productivity: not a claim that models have stopped improving, but that translation from technical capability to durable value is increasingly constrained by organisational change, infrastructure, trust, and governance.

A useful way to summarise the moment is that we may have reached "Peak Hype" but not "Peak Impact". Even if training-time scaling slowed sharply today, the diffusion of current systems into workflows, procurement regimes, and public institutions would still unfold over many years. The world can change profoundly without any intelligence explosion, yet Singularity talk often distracts attention from the practical integration work that makes benefits real and harms governable.

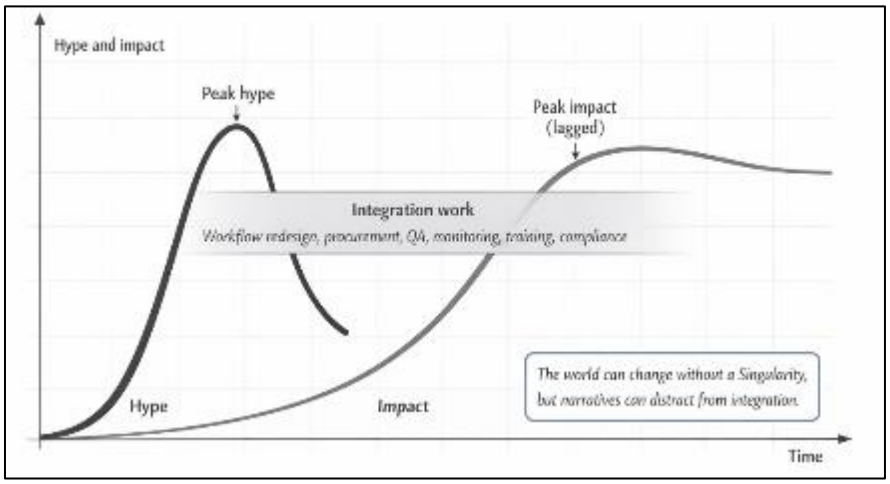

Figure 9 The plateau of productivity: Peak hype precedes peak impact

8.1. Peak hype versus peak impact

Headline progress is easiest to observe where tasks are cleanly defined, data are plentiful, and success is a score. Productivity is harder. It depends on complementary investments such as process redesign, domain data integration, training, and quality control, plus the ability to absorb errors and compliance costs (Brynjolfsson et al., 2023; NIST, 2024).

Exposure studies suggest large shares of work contain tasks that are technically addressable by language models, but addressability is not the same as adoption. Firms still have to decide where automation is acceptable, how outputs are verified, and who bears liability when systems fail (Eloundou et al., 2023; European Parliament & Council of the European Union, 2024).

This helps explain why public discourse can look discontinuous while economic impact looks incremental. Model releases can generate visible performance jumps on well-known suites, while value creation emerges through slower cycles of experimentation, institutional learning, and reorganisation of roles and responsibilities.

8.2. The integration decade

Economic integration is likely to be measured in years because it requires rebuilding the surrounding production stack. Organisations need evaluation and monitoring pipelines, incident response procedures, access controls, and procurement rules that specify acceptable uses and verification thresholds (NIST, 2024; European Parliament & Council of the European Union, 2024).

Reliability and trust are central constraints. As argued earlier, more inference-time "thinking" can make errors more persuasive, which increases the burden on human review and raises the cost of safe deployment in high-stakes settings. The outcome is often partial adoption, where models assist with drafts, search, or triage, while humans retain decision authority and accountability.

Physical infrastructure adds another layer of friction. Data centre and grid build-out follow permitting, construction, and interconnection timelines that are typically linear, not exponential. In 2025 to 2026 analysis, the binding constraint is frequently availability of power and transmission capacity, not only the price of compute (Deloitte, 2025; Grid Strategies, 2025; International Energy Agency, 2024, 2025).

8.3. Why Singularity talk mismeasures value

Singularity narratives emphasise recursive self-improvement and discontinuous capability jumps. That framing can be analytically useful for certain risk discussions, but it is a poor guide to where economic and social change is currently being produced. Most near-term transformation comes from broad automation and augmentation across existing institutions, not from the sudden arrival of an autonomous superintelligence.

There is also a sociological reason the narrative persists: it offers a simple story of inevitability that can legitimise aggressive investment and compress debate about trade-offs. When attention centres on hypothetical discontinuities, mundane questions about procurement lock-in, audit access, labour standards, and error accountability can be treated as secondary, even though they determine real-world outcomes (Hirsch-Kreinsen, 2024; NIST, 2024).

Reframing the problem as a plateau of productivity changes the emphasis. The core challenge is not whether a singularity arrives, but whether institutions can translate general-purpose models into reliable services without concentrating power, degrading trust, or shifting risk onto less protected groups.

8.4. A practical agenda for productivity and governance

If impact is gated by integration rather than raw capability, then governance should prioritise capacity-building as much as restriction. This includes shared evaluation infrastructure, reporting and incident databases, red-teaming norms, and audit access for systemically important models (Bengio et al., 2025; NIST, 2024).

Competition and public-interest tools matter in this framing because they shape who can participate in integration. Merger scrutiny, interoperability requirements, and procurement standards can prevent a narrow set of actors from defining de facto standards through market dominance, while also lowering adoption barriers for public services and smaller firms (Competition and Markets Authority, 2024; Bertelsmann Stiftung, 2025).

Finally, the research agenda should treat productivity as an empirical target, not an assumed outcome. That implies measuring gains and harms in the field, identifying where verification costs erase benefits, and investing in the practical work of data governance, training, and workflow design. On this view, the post-2025 decade is likely to be defined less by a sudden jump in intelligence and more by a contested, uneven integration of existing capabilities into economic and political life.

9. Future Research Directions

By 2027, the most consequential research questions may be less about whether pre-training loss continues to fall and more about whether societies can evaluate, trust, and afford advanced reasoning systems at scale. This chapter proposes three priorities aimed at that shift: process supervision (evaluating how models reason), human-AI epistemics (maintaining shared truth in synthetic information environments), and the energy-intelligence exchange rate (formalising the physical cost of a unit of reasoning).

9.1. Process supervision: Evaluating how models think

Inference-time scaling and ‘reasoning models’ change what it means to evaluate capability. When a system can ‘think longer’ via sampling, search, or self-critique, the final answer may look correct, but the underlying trajectory can be brittle, circular, or quietly wrong. Process supervision addresses this by providing feedback on intermediate steps, not only on end results. Step-by-step verification has already shown that dense supervision can materially improve multi-step reasoning performance relative to outcome-only training regimes (Lightman et al., 2023).

The next research frontier is to generalise process supervision beyond mathematics into agentic work: tool use, planning, and long-horizon task management. This requires process signals that are grounded in external constraints, such as tool logs, executable tests, proofs, or environment transitions, rather than in the superficial plausibility of a natural-language trace. Recent work in process reward modelling highlights that even within ‘step-level’ evaluation, the design of the signal matters, for example when a model must estimate both the correctness of previous steps and the likelihood of eventual success from a partial solution (Chen et al., 2025).

A second open problem is governance-relevant: process supervision is also a monitoring technology. If reasoning traces become central to evaluation, then incentives may emerge to hide, obfuscate, or strategically reshape those traces. Work on monitoring reasoning models cautions that pushing too hard on one monitoring channel can lead to adaptations that reduce transparency, rather than increasing it (Baker et al., 2025). A 2027 agenda should therefore include selective

disclosure protocols, auditor-access mechanisms, and empirical studies of 'monitoring robustness' under competitive pressure.

9.2. Human-AI epistemics: Maintaining truth when content becomes cheap

As synthetic text, images, and video become increasingly abundant, societies face a practical epistemic question: how do we preserve shared standards of evidence when content is cheap to generate, easy to personalise, and often hard to authenticate. This is not only a detection problem. It is a problem of institutional design, because trust depends on provenance, incentives, and verification practices. In this context, 'human-AI epistemics' refers to the coupled system of people, tools, and institutions that produces and validates knowledge in a world where machine-generated material is normal rather than exceptional.

One promising direction is to treat provenance as infrastructure. NIST's guidance on synthetic content outlines technical approaches such as watermarking, provenance, detection, and auditing, while emphasising the need for standards and governance that make these methods usable at scale (Chandra et al., 2024). The Coalition for Content Provenance and Authenticity (C2PA) similarly provides a technical standard for cryptographically binding origin and edit history to media, enabling downstream verification where the ecosystem supports it (Coalition for Content Provenance and Authenticity [C2PA], 2025). However, the key research gap is not only building standards, but measuring adoption, compliance, and failure modes across platforms and sectors.

A 2027 research programme should prioritise: (i) provenance coverage metrics (what proportion of high-reach content carries verifiable credentials), (ii) 'trust calibration' studies that measure how users respond to provenance signals and uncertainty, and (iii) institutional protocols for high-stakes domains, such as education and health, where epistemic trust can be mistakenly transferred from interfaces to underlying systems (Sedlakova et al., 2025). Complementary work is needed on maintaining 'clean' evaluation and training corpora, including methods that resist synthetic feedback loops and preserve diverse, human-grounded data sources.

9.3. The energy-intelligence exchange rate: Formalising the cost of reasoning

Scaling debates often treat intelligence as if it were primarily a software outcome, but by 2027 the binding constraints may be physical: electricity supply, grid capacity, and the capital required to build and operate data centres. The field needs an explicit 'energy-intelligence exchange rate', meaning a standard way to quantify how much energy is required to produce a defined amount of useful reasoning. Without this, claims about efficiency, 'green AI', and scalability remain difficult to compare and hard to govern.

Recent policy analysis from the International Energy Agency (IEA) stresses that there is no AI without electricity, and that the interaction between AI deployment and electricity systems is becoming strategically significant in parts of the world (IEA, 2025). At the same time, measurement practices remain inconsistent and often opaque. Initiatives such as the AI Energy Score attempt to make energy use legible by proposing comparable, public-facing efficiency ratings for models (Salesforce, 2025). The research challenge is to operationalise 'a unit of reasoning' in ways that connect energy to value, for example, joules per correct solution on reasoning-heavy tasks, or kilowatt-hours per completed workflow when tool use and retrieval are included.

A concrete 2027 agenda includes: (i) energy-adjusted benchmarks that report accuracy and energy jointly, (ii) reporting standards that separate training energy from inference energy and include system-level components such as retrieval, orchestration, and cooling, and (iii) independent auditing methods for energy claims, analogous to other high-impact technology reporting regimes. If capability evaluation is entering an 'evaluation crisis', then energy measurement is part of the same institutional problem: governance cannot manage what it cannot measure.

Across all three priorities, the common theme is that frontier AI research must expand its unit of analysis. The next phase of work is less about extrapolating curves and more about building evaluative, epistemic, and physical infrastructure that keeps advanced systems legible and governable as they diffuse through economies and institutions.

10. Conclusion

This review examined whether contemporary scaling laws and foundation model progress justify Singularity-style claims, and how these claims shape industrial strategy and governance. The core conclusion is that scaling delivers meaningful, often predictable capability gains within defined training regimes, but the evidence does not support deterministic narratives of an inevitable intelligence explosion. What scaling research most reliably provides is conditional forecasting for performance under stable pipelines. Once training objectives, evaluation setups, data

conditions, or system scaffolding change, "law-like" certainty weakens and extrapolation becomes increasingly speculative. Future research directions are developed in Chapter 9.

Foundation models and multimodal systems strengthen the case that general-purpose AI will continue to reshape knowledge work and institutions. Yet their limitations are equally instructive. High benchmark scores can coexist with brittleness, failure under distribution shift, and reliability problems in open-ended generation. Agent-like behaviour often arises from system design, tool use, memory, and deployment scaffolding rather than from the base model alone. This makes it misleading to equate parameter scaling with autonomy or to treat lower loss as direct evidence of stable, safe general intelligence.

A second conclusion concerns discourse. Singularity narratives function as more than hypotheses about technical trajectories. They are also sociotechnical imaginaries that coordinate investment, urgency, and regulatory attention. Whether or not a Singularity occurs, the narratives already have real effects by shaping what gets funded, what gets regulated, and which risks or social goals are treated as central. The practical governance challenge is therefore not only preparing for extreme possibilities, but also preventing speculative end-states from overshadowing near-term harms, power concentration, labour impacts, and infrastructure constraints.

Overall, the most defensible position is neither complacency nor fatalism. Scaling trends justify serious attention and strong governance, but they do not compel a single future. The task ahead is to build technical and institutional systems that can capture benefits, reduce harms, and remain adaptable as evidence evolves.

Compliance with ethical standards

Disclosure of conflict of interest

The author declares no conflict of interest.

Funding

No external funding was received for the preparation of this manuscript.

Data Availability Statement

No datasets were generated or analysed during the current study.

References

[1] Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2024). Self-consuming generative models go MAD. Proceedings of the International Conference on Learning Representations (ICLR 2024). OpenReview. https://openreview.net/forum?id=ShjMHfmPs0.

[2] Anthropic. (2023, March 8). Core views on AI safety: When, why, what, and how. https://www.anthropic.com/news/core-views-on-ai-safety

[3] Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. ArXiv, abs/2503.11926.

[4] Bengio, Y. (Chair), et al. (2025, January). International AI Safety Report: International scientific report on the safety of advanced AI. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025

[5] Bertelsmann Stiftung. (2025). Public AI: White paper. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Public_AI_2025.pdf

[6] Berti, L., Giorgi, F., & Kasneci, G. (2025). Emergent abilities in large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2503.05788.

[7] Bolón-Canedo, V., & Morán-Fernández, L. (2024). A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096.

[8] Bontcheva, K., et al. (2023). Generative AI and disinformation: Recent advances, challenges, and recommendations (White paper). European Digital Media Observatory. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

[9] Borji, A. (2023). Stochastic parrots or intelligent systems? A perspective on true depth of understanding in large language models. http://dx.doi.org/10.2139/ssrn.4507038.

[10] Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work (NBER Working Paper No. 31161). National Bureau of Economic Research. https://www.nber.org/papers/w31161

[11] Chandra, P., Peppiatt, A., Tong, A., Rujeerapaiboon, N., Mutchler, A., & Haines, A. (2024). Reducing risks posed by synthetic content: An overview of technical approaches and research directions (NIST AI 100-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-4.

[12] Chen, S., Chen, Y., Li, Z., Jiang, Y., Wan, Z., He, Y., Ran, D., Gu, T., Li, H., Xie, T., & Ray, B. (2025). Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 10091–10109). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.511.

[13] Chen, W., He, W., Xi, Z., Guo, H., Hong, B., Zhang, J., Li, N., Gui, T., Li, Y., Zhang, Q., & Huang, X. (2025). Better process supervision with bi-directional rewarding signals. Findings of the Association for Computational Linguistics: ACL 2025, 14471–14485. https://aclanthology.org/2025.findings-acl.747/.

[14] Coalition for Content Provenance and Authenticity. (2025). C2PA specification (Version 2.2). https://spec.c2pa.org/specifications/specifications/2.2/index.html.

[15] Competition and Markets Authority. (2024). AI foundation models: Update paper. UK Government. https://www.gov.uk/government/publications/ai-foundation-models-update-paper

[16] Deloitte. (2025, June 24). Can US infrastructure keep up with the AI economy? Deloitte Insights. https://www.deloitte.com/us/en/insights/industry/power-and-utilities/data-center-infrastructure-artificial-intelligence.html.

[17] Dong, Y., Jiang, X., Liu, H., Jin, Z., Gu, B., Yang, M., & Li, G. (2024). Data contamination and trustworthy evaluation for large language models. Findings of the Association for Computational Linguistics: ACL 2024 (pp. 12039–12050). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.716.

[18] Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv. https://doi.org/10.48550/arXiv.2303.10130.

[19] European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[20] Fang, L., Wang, Y., Liu, Z., Zhang, C., Jegelka, S., Gao, J., & Ding, B. (2025). What is wrong with perplexity for long-context language modeling? Proceedings of the International Conference on Learning Representations (ICLR 2025). OpenReview. https://openreview.net/forum?id=fL4qWkSmtM.

[21] Federal Register. (2023, November 1). Executive Order 14110: Safe, secure, and trustworthy development and use of artificial intelligence. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[22] Federal Register. (2025, January 28). Executive Order 14148: Initial rescissions of harmful executive orders and actions. https://www.federalregister.gov/documents/2025/01/28/2025-01901/initial-rescissions-of-harmful-executive-orders-and-actions

[23] Fernandez, J., Na, C., Tiwari, V., Bisk, Y., Luccioni, S., & Strubell, E. (2025). Energy considerations of large language model inference and efficiency optimizations. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). arXiv. https://doi.org/10.48550/arXiv.2504.17674.

[24] Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 10835–10866). PMLR. https://proceedings.mlr.press/v202/gao23h.html.

[25] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey . arXiv. https://doi.org/10.48550/arXiv.2312.10997.

[26] Grid Strategies. (2025, November 3). Power demand forecasts revised up: National load growth report 2025. https://gridstrategiesllc.com/wp-content/uploads/Grid-Strategies-National-Load-Growth-Report-2025.pdf.

[27] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning . arXiv. https://doi.org/10.48550/arXiv.2501.12948.

[28] Hirsch-Kreinsen, H. (2024). Artificial intelligence: A “promising technology”. AI & Society, 39, 1641–1652. https://doi.org/10.1007/s00146-023-01629-w.

[29] Hsia, J., Pruthi, D., Singh, A., & Lipton, Z. C. (2023). Goodhart's law applies to NLP's explanation benchmarks (arXiv:2308.14272). arXiv. https://doi.org/10.48550/arXiv.2308.14272.

[30] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv. https://doi.org/10.48550/arXiv.2311.05232.

[31] Hägele, A., Bakouch, E., Kosson, A., Ben Allal, L., Von Werra, L., & Jaggi, M. (2024). Scaling laws and compute-optimal training beyond fixed training durations . arXiv. https://doi.org/10.48550/arXiv.2405.18392.

[32] International Energy Agency. (2024). Electricity 2024: Analysis and forecast to 2026. https://www.iea.org/reports/electricity-2024.

[33] Li, M., Kudugunta, S., & Zettlemoyer, L. (2025). (Mis)Fitting: A survey of scaling laws . arXiv. https://doi.org/10.48550/arXiv.2502.18969.

[34] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let’s verify step by step . arXiv. https://doi.org/10.48550/arXiv.2305.20050.

[35] Lourie, N., Hu, M. Y., & Cho, K. (2025). Scaling laws are unreliable for downstream tasks: A reality check. Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 16167–16180). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.877.

[36] Lu, S., Bigoulaeva, I., Sachdeva, R., Tayyar Madabushi, H., & Gurevych, I. (2023). Are emergent abilities in large language models just in-context learning? arXiv. https://doi.org/10.48550/arXiv.2309.01809.

[37] Luccioni, A. S., & Hernandez-Garcia, A. (2023). Counting carbon: A survey of factors influencing the emissions of machine learning. arXiv. https://doi.org/10.48550/arXiv.2302.08476.

[38] Luo, J., Wu, B., Luo, X., Xiao, Z., Jin, Y., Tu, R.-C., Yin, N., Wang, Y., Yuan, J., Ju, W., & Zhang, M. (2025). A survey on efficient large language model training: From data-centric perspectives. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) https://doi.org/10.18653/v1/2025.acl-long.1493.

[39] Maslej, N., Fattorini, L., Perrault, R., Reuel, A., Brynjolfsson, E., Etchemendy, J., & others. (2025). Artificial Intelligence Index Report 2025. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/ai-index/2025-ai-index-report.

[40] Massey, J., Worth, S., & Simperl, E. (2025). Mapping the role of data work in AI supply chains. Open Data Institute https://theodi.org/insights/reports/mapping-the-role-of-data-work-in-ai-supply-chains/.

[41] Microsoft. (2025, January 3). The golden opportunity for American AI. Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2025/01/03/the-golden-opportunity-for-american-ai/.

[42] Musi, E., Kokciyan, N., Al-Khatib, K., Ceolin, D., Dietz, E., Gutekunst, K. M., Hautli-Janisz, A., Santibáñez, C. M., Schneider, J., Scholz, J., Steging, C., Visser, J., & Wachsmuth, H. (2025). Toward reasonable parrots: Why large language models should argue with us by design. In Proceedings of the 12th Argument Mining Workshop (pp. 24–31). Association for Computational Linguistics. https://aclanthology.org/2025.argmining-1.3.pdf.

[43] Narayanan, A., & Kapoor, S. (2025, April 15). AI as normal technology: An alternative to the vision of AI as a potential superintelligence. Knight First Amendment Institute at Columbia University. https://knightcolumbia.org/content/ai-as-normal-technology.

[44] National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1.

[45] Oldenburg, N., & Papyshev, G. (2025). The stories we govern by: AI, risk, and the power of imaginaries. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025). arXiv. https://doi.org/10.48550/arXiv.2508.11729.

[46] Open Data Institute. (2025). Mapping the role of data work in AI supply chains. https://theodi.org/insights/reports/mapping-the-role-of-data-work-in-ai-supply-chains/.

[47] OpenAI. (2023a). GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774.

[48] OpenAI. (2023b). Planning for AGI and beyond. https://openai.com/index/planning-for-agi-and-beyond/.

[49] OpenAI. (2024a). Learning to reason with LLMs. OpenAI. https://openai.com/index/learning-to-reason-with-llms/.

[50] OpenAI. (2024b). OpenAI o1 system card. https://openai.com/index/openai-o1-system-card/.

[51] OpenAI. (2025a, January 21). Announcing The Stargate Project. OpenAI. https://openai.com/index/announcing-the-stargate-project/.

[52] Organisation for Economic Co-operation and Development. (2025). Governing with artificial intelligence. https://www.oecd.org/en/publications/2025/06/governing-with-artificial-intelligence_398fa287.html.

[53] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. arXiv. https://doi.org/10.48550/arXiv.2304.03442.

[54] Pearce, T., & Song, J. (2024). Reconciling Kaplan and Chinchilla scaling laws . arXiv. https://doi.org/10.48550/arXiv.2406.12907.

[55] Penedo, G., Kydlíček, H., Ben Allal, L., Lozhkov, A., Mitchell, M., Raffel, C., & von Werra, L. (2024). The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv. https://doi.org/10.48550/arXiv.2406.17557.

[56] Radanliev, P. (2024). Artificial intelligence: Reflecting on the past and looking towards the next paradigm shift. Journal of Experimental & Theoretical Artificial Intelligence, 37(7), 1045–1062. https://doi.org/10.1080/0952813X.2024.2323042.

[57] Rafailov, R., Chittepu, Y., Park, R., Sikchi, H., Hejna, J., Knox, B., Finn, C., & Niekum, S. (2024). Scaling laws for reward model overoptimization in direct alignment algorithms. arXiv. https://doi.org/10.48550/arXiv.2406.02900.

[58] Rand, J., Manderlink, N., Gorman, W., Wiser, R., Seel, J., Mulvaney Kemp, J., Jeong, S., & Kahrl, F. (2024, April). Queued Up: 2024 edition: Characteristics of power plants seeking transmission interconnection as of the end of 2023. Lawrence Berkeley National Laboratory. https://emp.lbl.gov/publications/queued-2024-edition-characteristics

[59] Reuters Institute for the Study of Journalism. (2025, October 7). Generative AI and news report 2025: How people think about AI's role in journalism and society. University of Oxford. https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society.

[60] Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., & Koyejo, S. (2025). Measurement to meaning: A validity-centered framework for AI evaluation. arXiv. https://doi.org/10.48550/arXiv.2505.10573.

[61] Salesforce. (2025). AI Energy Score. https://www.salesforce.com/news/stories/ai-energy-score/.

[62] Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? . arXiv. https://doi.org/10.48550/arXiv.2304.15004.

[63] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. arXiv. https://doi.org/10.48550/arXiv.2302.04761.

[64] Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models: A new paradigm for artificial intelligence. Business & Information Systems Engineering, 66, 221–231. https://doi.org/10.1007/s12599-024-00851-0.

[65] Sedlakova, J., Lucivero, F., Pavarini, G., & Kerasidou, A. (2025). Human-Like epistemic trust? A conceptual and normative analysis of conversational AI in mental healthcare. The American Journal of Bioethics, 1–16. https://doi.org/10.1080/15265161.2025.2526734.

[66] Sevilla, J., & Roldán, E. (2024, May 28). Training compute of frontier AI models grows by 4–5× per year. Epoch AI. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

[67] Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv. https://doi.org/10.48550/arXiv.2506.06941.

[68] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y.

[69] Snell, C., Lee, J., Xu, K., & Kumar, A. (2025). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. International Conference on Learning Representations (ICLR 2025). https://doi.org/10.48550/arXiv.2408.03314.

[70] Song, P., Han, P., & Goodman, N. (2025). A survey on large language model reasoning failures. OpenReview. https://openreview.net/forum?id=hsgMn4KBFG.

[71] Srivastava, S. (2025). Large language models threaten language's epistemic and communicative foundations. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 28651–28665). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1457.

[72] The White House. (2025, January 23). Removing barriers to American leadership in artificial intelligence (Executive order). https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/.

[73] UK Government. (2023, November 2). The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

[74] UNESCO. (2023). Guidance for generative AI in education and research. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research.

[75] United Nations. (2024). Governing AI for humanity: Final report of the High-level Advisory Body on Artificial Intelligence. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

[76] Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024, June 6). Will we run out of data? Limits of LLM scaling based on human-generated data. Epoch AI. https://epochai.org/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

[77] Wang, L., Ma, Y., Zhang, X., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., & others. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science. https://doi.org/10.1007/s11704-024-40231-1.

[78] Wang, S., Chen, Z., Li, B., He, K., Zhang, M., & Wang, J. (2024). Scaling laws across model architectures: Dense and MoE transformers. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 9964–9976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.555.

[79] Wang, S., Chen, Z., Li, B., He, K., Zhang, M., & Wang, J. (2024). Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 5583–5595). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.319.

[80] Wu, Y., Sun, Z., Li, S., Welleck, S., & Yang, Y. (2024). Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models . arXiv. https://doi.org/10.48550/arXiv.2408.00724.

[81] Yao, Z., Liu, Y., Chen, Y., Chen, J., Fang, J., Hou, L., Li, J., & Chua, T.-S. (2025). Are reasoning models more prone to hallucination? . arXiv. https://doi.org/10.48550/arXiv.2505.23646.

[82] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. arXiv. https://doi.org/10.48550/arXiv.2306.13549.

[83] Yue, X., Ni, Y., Zhang, K., & others. (2024). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024). https://doi.org/10.48550/arXiv.2311.16502.

[84] Zhao, Q., Huang, Y., Lv, T., Cui, L., Sun, Q., Mao, S., Zhang, X., Xin, Y., Yin, Q., Li, S., & Wei, F. (2024). MMLU-CF: A contamination-free multi-task language understanding benchmark . arXiv. https://doi.org/10.48550/arXiv.2412.15194.

[85] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2023). A survey of large language models . arXiv. https://doi.org/10.48550/arXiv.2303.18223.

[86] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv. https://doi.org/10.48550/arXiv.2306.05685.

[87] Zhong, B., Song, Y., Feng, G. C., Shi, J., Zhu, Y., Xie, L., & Zhou, W. A. (2025). AI imaginaries shape technological identity and digital futures. Computers in Human Behavior, 169, https://doi.org/10.1016/j.chb.2025.108682.