A Bayesian framework for forecasting and scenario analysis of AI task automation under scaling constraints

Xiaoliang Luo[1, *, †], Hannan Saddiq[2, *], Rachel Asquith[3], Srish Bhasin[4], Jamie McGraw[5], and Marc Warner[6]

[1,3,4,5,6] Faculty AI, [2] work done while at Faculty AI, *Equal contribution, †Correspondence to: xiaoliang.luo@faculty.ai,

Abstract

The rate of AI progress has become one of the most critical questions of our time. It has become a load-bearing assumption across a remarkable breadth of domains — healthcare, education, defence, geopolitics, finance, and employment, to name a few. Yet while scaling laws reliably predict AI performance in abstract benchmarks, forecasting the automation of real-world labor remains speculative due to a lack of integrated models that account for physical and algorithmic limits. We present a hierarchical Bayesian framework that causally links primary resource constraints—electrical power availability, infrastructure efficiency, and algorithmic progress—to success probabilities on professional software engineering tasks. By jointly modeling previously disconnected data on power projections, training compute, algorithmic efficiency, and real-world task performance, we propagate uncertainty across the entire causal chain from physical infrastructure to task-level outcomes. Unlike naive extrapolations that imply continued exponential growth in AI capability, our framework enables systematic scenario analysis—placing

competing views about how key constraints evolve on equal footing and directly comparing their implications across the pace of power infrastructure build-out, the share of resources directed toward AI, and the rate at which algorithmic efficiency gains saturate. As empirical understanding of AI development matures, new estimates of power availability, revised models of algorithmic progress, or entirely new constraints can each be incorporated—making this a living tool for researchers, policymakers, and industry stakeholders to systematically test their own assumptions and translate them into actionable intelligence for economic and policy planning.

1 Introduction

Scaling laws are quantitative relationships between system size and performance that have emerged as fundamental principles organizing complex systems from biology to artificial intelligence. Allometric scaling in organisms predicts metabolic requirements across species (West et al., 1997). Power-law connectivity in networks characterizes diverse technological and natural systems (Barabási and Albert, 1999). In machine learning, the discovery that model performance scales predictably with compute, parameters, and data—following power laws across orders of magnitude—has fundamentally transformed how AI systems are developed and deployed, enabling researchers to forecast computational demands and prioritize investments (Kaplan et al., 2020; Hoffmann et al., 2022; Snell et al., 2024).

More recently, these scaling principles have been extended to a question with immediate societal and economic implications: which real-world tasks will AI systems be capable of automating, and when (Kwa et al., 2025; Erdil et al., 2025; Patwardhan et al., 2025)? Understanding the trajectory of AI task automation capabilities is essential for multiple stakeholders—investors assessing economic disruption, policymakers designing governance frameworks, and workers across diverse professions preparing for labor market transitions (Eloun-

dou et al., 2023; Patwardhan et al., 2025). Forecasting AI automation capabilities longitudinally—predicting not just whether tasks can be automated, but when—translates abstract scaling laws into actionable intelligence for resource planning, workforce development, and policy design across government, industry, and civil society.

One seminal framework for longitudinal forecasting was introduced by Kwa et al. (2025) from METR (Model Evaluation & Threat Research), which identified that the "50%-task-completion time horizon"—the maximum task duration a model can handle with 50% success—has been doubling approximately every seven months. This methodology benchmarks AI against human professional completion times, providing a continuous metric for capability growth. However, naive temporal extrapolation of model capability neglects to account for physical, economic and algorithmic boundaries of scaling. Recent analyses (You et al., 2025) highlight that the staggering power requirements of frontier training may soon encounter structural limits. Moreover, Edelman and Ho (2025) demonstrate that compute scaling will slow due to increasing lead times for datacenter construction and chip manufacturing. Building on similar insight, Whitfill et al. (2025) extended the METR framework by establishing a power-law relationship between compute and AI agent time horizons, incorporating algorithmic progress through an ideas production function dependent on experimental compute. They model compute slowdowns by translating OpenAI's projected R&D expenditures into FLOPs trajectories, thereby embedding economic and infrastructure constraints into capability forecasts through the financial spending ceiling. Despite this progress, a unified model that coherently integrates infrastructure constraints with specific task characteristics—such as complexity and duration—remains elusive.

In this contribution, we propose a unified hierarchical framework for forecasting AI automation that jointly models the causal relationship from primary resource constraints—electrical power availability, infrastructure efficiency, and algorithmic progress—to downstream task per-

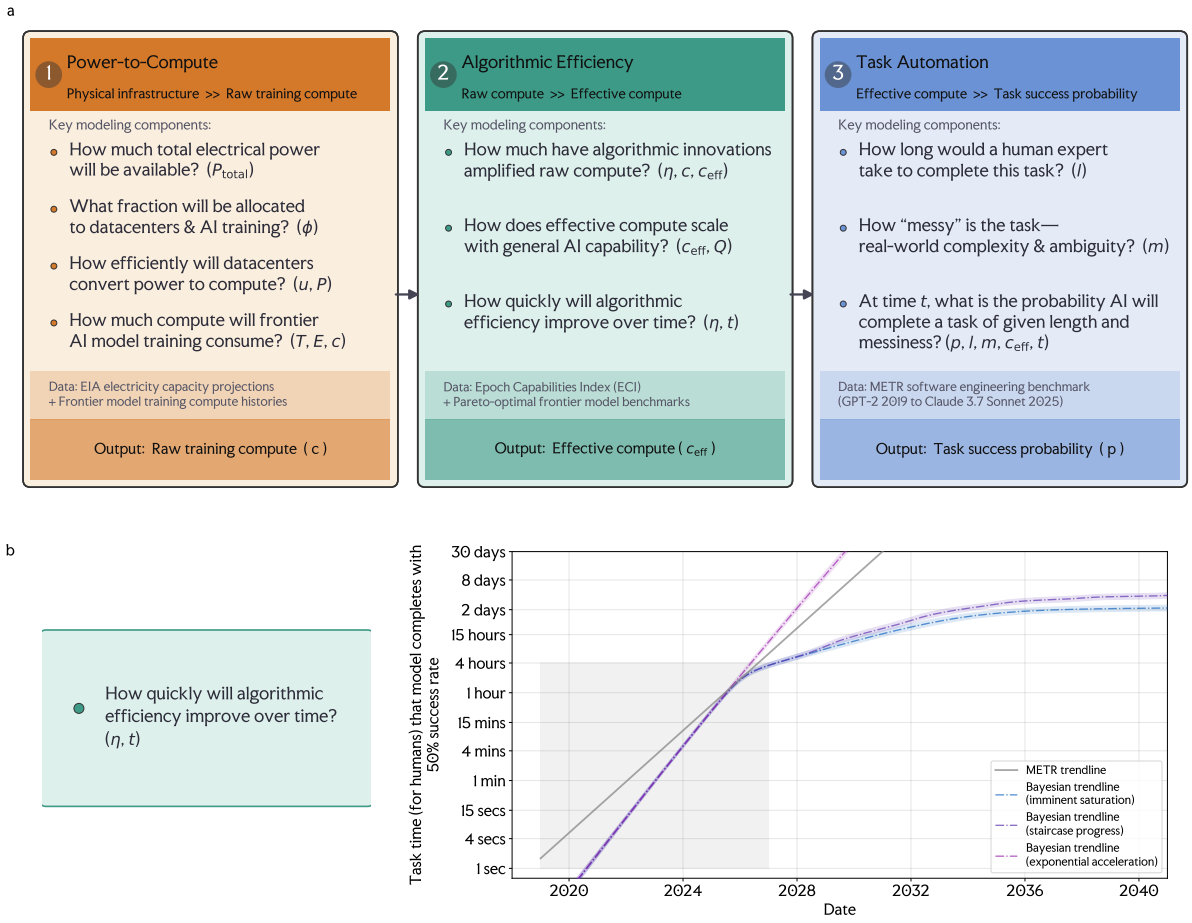

Figure 1: An overview of the hierarchical forecasting framework. (a) The framework consists of three interconnected layers. The **power-to-compute layer** models how time (t) and U.S. total power capacity (P_{total}) flow through allocation fractions (ϕ) and efficiency factor—power usage effectiveness (u)—to determine realized power (P), and through training duration (T) to determine training energy (E) and compute (c). The **algorithmic efficiency layer** converts raw compute (c) into effective compute (c_{eff}) via time-dependent algorithmic improvements (η), which relates to general capability (Q) measured by the Epoch Capabilities Index. The **task automation layer** maps effective compute to task-specific outcomes conditioned on task length (l) and messiness (m), producing success probabilities (p) for METR software engineering tasks. Arrows represent causal dependencies flowing from infrastructure constraints through computational resources and algorithmic efficiencies to task success. (b) Forecasted AI task automation capabilities under three algorithmic-efficiency scenarios for tasks at average task complexity from the METR benchmark: imminent saturation (sigmoid; blue dash-dot), staircase progress (sum of shifted sigmoids; purple dash-dot) and exponential acceleration (violet dash-dot). See results for details.

formance (Figure 1a). Where prior approaches treat these factors in isolation, our framework yields a coherent, internally consistent picture of AI progress that is grounded in physical realities and how different constraints interact.

Our framework is built on three interconnected causal components, illustrated in Figure 1a. First, the power-to-compute layer translates available electrical power into training compute by modeling U.S. total power capacity (P_{total}), allocation fractions to AI training, and time-varying factors in both datacenter infrastructure—captured by power usage effectiveness (u) and datacenter resource allocation for training frontier models (ϕ). This produces the realized power (P) along with the training duration (T) constrains the training compute (c) available to frontier AI systems. We inform this component using models on the time-compute or time-capability frontier: those that either represented the highest training compute at their release or achieved state-of-the-art performance on practical software engineering tasks from the METR benchmark (Kwa et al., 2025)—a suite of 170 real-world tasks spanning from one-minute coding problems to multi-day machine learning research projects (All experiments use version time-horizon-1-0; note that the benchmark is under active development).

Second, the algorithmic efficiency layer (Figure 1a) captures how algorithmic innovations convert raw compute (c) into "effective compute" (c_{eff}) through a time-varying efficiency factor (η). Effective compute is then related to general AI capability using the Epoch Capabilities Index (Q)—a composite measure aggregating performance across diverse benchmarks into a single capability scale (EpochAI, 2025). This component is informed by models on the time-capability-compute frontier: systems that achieved high capabilities relative to their training compute, representing algorithmic or architectural innovations.

Third, the task automation layer maps effective compute to success probabilities (p) on specific tasks, ultimately predicting task outcomes. This component accounts for task-specific dimensions including completion time horizons (l—the time a human professional would need)

and task complexity or "messiness" (m)—a quantified measure of factors like resource constraints, dynamic requirements, and implicit specifications that make real-world tasks harder than sanitized benchmarks (Kwa et al., 2025). We model this using the METR task automation data, where frontier language models from GPT-2 (2019) to Claude 3.7 Sonnet (2025) were evaluated on their ability to complete these practical software engineering challenges (Kwa et al., 2025).

We learn this system via hierarchical Bayesian inference, enabling us to jointly inform our estimates with previously disconnected datasets—including U.S. electrical capacity projections (EIA, 2024), frontier model training histories (You et al., 2025), algorithmic efficiency benchmarks from Pareto-optimal models (EpochAI, 2025), and empirical performance on software engineering tasks (Kwa et al., 2025). This joint inference approach propagates uncertainty across the entire causal chain—from power allocation through compute scaling to task-specific outcomes—while simultaneously estimating functional relationships between compute and capability from frontier model data and learning algorithmic efficiency patterns from models that achieve high performance with minimal resources. By partitioning our training data into complementary frontiers, we can separately identify the contributions of scale versus algorithmic innovation. For more technical details, please refer to Methods.

Crucially, we do not claim to offer immutable predictions; rather, we provide a systematic, transparent lens for forecasting that reconciles abstract scaling laws with the physical realities of infrastructure. The framework's modular design also enables counterfactual scenario analysis, allowing different constraint assumptions to be swapped in and their implications compared. For instance, rather than committing to a definitive trajectory of algorithmic improvement, our framework can compare interchangeable parametric forms, such as scaled sigmoid, exponential, or sum-of-sigmoid, that may fit historical observations equally well yet produce divergent forecasts. As a preview of our key results, Figure 1b contrasts exponential extrapolation of

the METR framework with our Bayesian forecasts under different algorithmic improvement scenarios. Where METR implies continued exponential growth in task-completion horizons, our framework yields markedly different trajectories depending on how physical and algorithmic constraints bind. Beyond algorithmic scenarios, the framework can similarly accommodate variation in total power growth and in the fraction of power directed toward AI training infrastructure—trajectories we unpack in Section 2.

As better estimates of power availability, compute efficiency, and algorithmic progress become available, the framework can be readily updated. We view this work as establishing methodological groundwork for more sophisticated forecasting as empirical understanding of AI development matures, and as a systematic tool for scenario analysis—enabling researchers, policymakers, and industry stakeholders to explore how different assumptions about infrastructure investment, algorithmic breakthroughs, or resource allocation might alter automation timelines.

2 Results

Our integrated framework for forecasting AI task automation success is constructed around key causal constraints: electrical power availability, the efficiency of converting power to computation, and the rate of algorithmic progress. The model is a Bayesian hierarchical structure where each component is informed by specialized datasets and can be independently swapped out for counterfactual analysis under different development assumptions. This section is divided in two: we first fix a reference scenario and trace the learning of each model component in turn to derive AI growth projections, then examine how forecasts change under alternative assumptions about power supply, the fraction of data center power invested in AI training, and algorithmic improvement.

2.1 Modeling AI Development Under Scaling Constraints

2.1.1 The Power-Compute Frontier

A fundamental constraint on the scaling of frontier AI is the availability of electrical power. We first model the realized power draw for training AI systems by starting with the total U.S. electrical power capacity (using data from the Energy Information Administration; EIA 2024) and applying a series of allocation fractions: the share consumed by datacenters, the share within datacenters dedicated to AI, and the portion of AI power used for training, all divided among competing hyperscale companies (see Methods).

The resulting power bound is then modulated by two time-varying factors: **Power Usage Effectiveness (PUE)** $u(t)$, which is the inverse of datacenter infrastructure efficiency which decreases over time as datacenter overhead diminishes (Equation 5). **Utilization Fraction** $\phi(t)$, which captures the fraction of total datacenter power dedicated to training a frontier model–a utilization allocation fraction which increases over time (Equation 6). The combination of these factors determines the fraction of power bound that is realized for computation (Equation 7).

Figure 2a shows the posterior estimates of the realized power draw available for training individual frontier models from 2018–2040. The model successfully captures the rapid, historical growth of training power (black points, sourced from Epoch AI's analysis of frontier models; You et al. 2025), with the median fit (dark orange line) tracking historical data across orders of magnitude. Importantly, while a naive exponential extrapolation (dotted purple line) forecasts continued, unrestrained growth, our physically-constrained model projects a significant slowing of power growth beginning mid-2030, as the increasing demand for AI training power begins to reach structural limits imposed by total US capacity and realistic efficiency gains. The reference scenario assumes total power supply follows the EIA baseline growth trajectory (EIA, 2024) and that the fraction of data center power dedicated to AI training remains fixed at its estimated current level (December 2025; details see Supplementary Information).

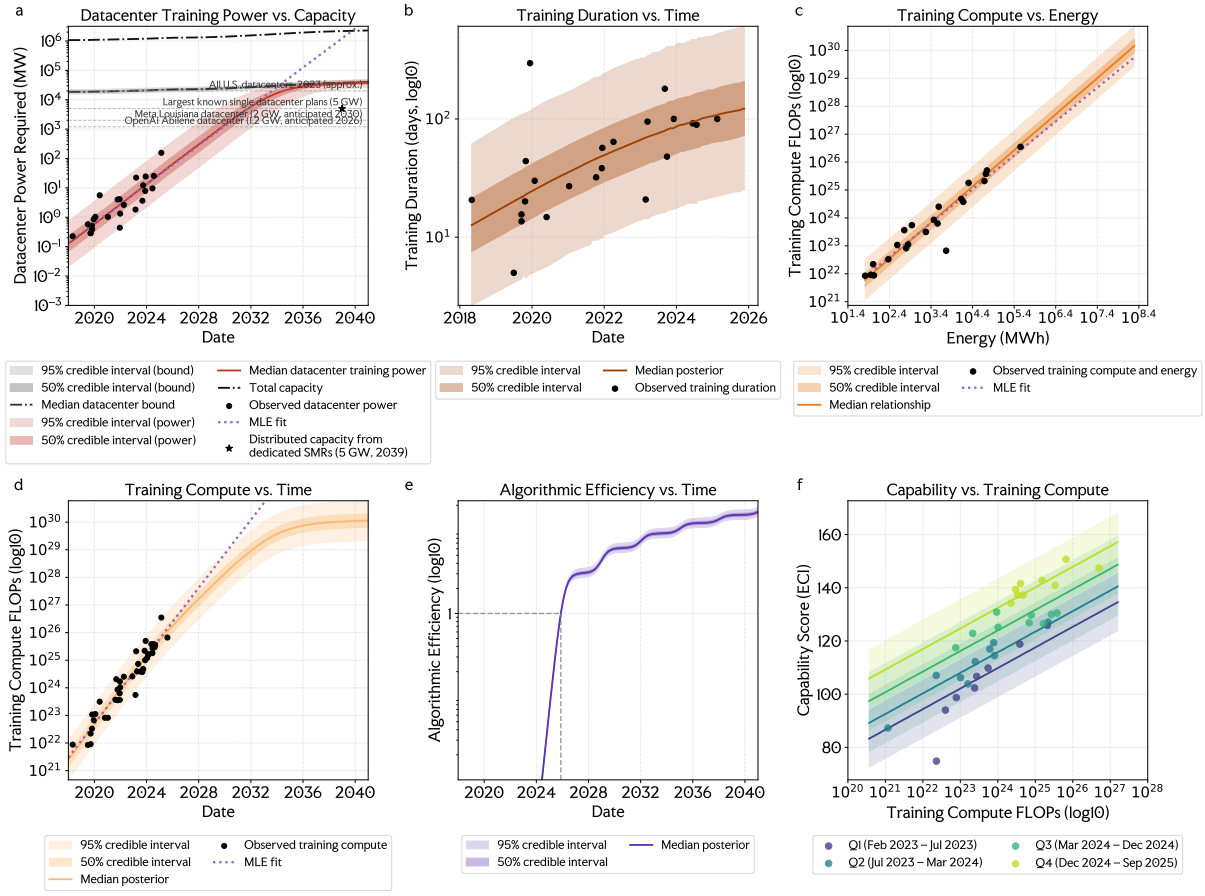

Figure 2: AI training compute under constraints. (a) Posterior estimates of datacenter training power available for frontier models relative to total U.S. capacity. The median fit (dark orange) tracks historical data (black points; You et al. 2025) but projects a slowdown mid-2030s as demand approaches structural limits; dotted purple line shows an unconstrained maximum likelihood extrapolation. (b) Training duration over time, with the median posterior increasing sigmoidally. (c) Log-linear relationship between realized training energy and training compute for frontier models. (d) Projected training compute derived from power constraints in (a), with growth slowing considerably in the mid-2030s. (e) Algorithmic efficiency modeled as a sum of shifted sigmoids curve, normalized to 1 at the latest observation. (f) Capability score (ECI; EpochAI 2025) as a function of training compute, with upward shifts across time quartiles reflecting algorithmic efficiency gains. Shaded bands represent 50% and 95% credible intervals throughout.

To convert constrained power to training compute, we first calculate energy expenditure by multiplying realized power by training duration. We then transform energy to training compute via a log-linear relationship (Equation 13). Training duration for individual models is modeled as a time-varying fraction of an upper bound, evolving sigmoidally according to historical data from You et al. (2025) (Section 4.3.4; Figure 2b). The energy-to-compute transformation (Figure 2c) is then fitted using frontier models from You et al. (2025)—systems representing maximum compute scale or capability at the time of release (see Methods).

The fitted relationship (Figure 2c) captures how historical frontier models have converted energy into compute, providing the basis for projecting future training compute availability under power constraints in Figure 2d. These compute estimates then serve as inputs to the task automation model, linking infrastructure limitations to AI capability forecasts.

2.1.2 Algorithmic Efficiency and Effective Compute

Beyond raw hardware scaling, AI progress is driven by algorithmic innovations that increase the efficiency of invested compute. We model this progress by defining **Algorithmic Efficiency** $\eta(t)$) as a parametric function over time (Section 4.4). For this reference scenario, algorithmic improvement is modeled as a sum of shifted sigmoids, reflecting the intuition that future progress will likely arrive in waves—punctuated by periodic breakthroughs and paradigm shifts, interspersed with periods of slower or saturated improvement. This trajectory is illustrative rather than prescriptive; we discuss alternative forms in the following sections and in the discussion. Efficiency is normalized to 1 at the latest observation (Figure 2e).

This efficiency factor converts raw training compute into the more relevant **Effective Compute**, which then serves as one of the key predictors for model capability. This component was informed by models on the time-capability-compute Pareto frontier (see Methods)—models that achieved a high capability score with a minimal amount of training compute, thus representing

efficient use of resources through innovation. We measure capability using the Epoch Capabilities Index (ECI; EpochAI 2025), a composite benchmark score. We model the relationship between capability and effective compute as a linear scaling law in log-space (Equation 22; model fit see Figure 2f).

2.1.3 Task Automation Success

Finally, we link the forecast for effective compute to the task automation objective–the probability of an LLM successfully completing a real-world task is modeled using a logistic regression framework (Equation 1), conditional on: **Task Length**, the time a human professional needs for completion; **Task Messiness**, which is a proxy for task complexity, based on real-world factors like resource constraints or dynamic environments; and **Effective Compute**, the effective training compute used to train the LLM performing the task. We model the task outcome as a Bernoulli distribution (completed or not). This component was trained using data from Kwa et al. (2025), a benchmark of 170 real-world software engineering tasks evaluated on a range of frontier LLMs. The trained model demonstrates performance of accuracy: 0.877, precision: 0.896, recall: 0.871, with calibration alignment across probability ranges and posterior predictive checks showing adequate capture of the underlying data-generating process (PPC p-value: 0.964; details in Supplementary Information).

2.1.4 Forecasting Time Horizons and Success Frontiers

To gain a better view of the potential trajectory of AI automation, our framework enables us to draw out the complete "success frontier" which illustrates tasks of varying complexities, lengths and their projected success rate being completed by AI over time. Under the reference scenario, Figure 3a–f illustrate the how such "success frontier" would shift over time: in 2022, models were largely limited to simple, low-messiness tasks lasting only a few minutes; by 2030, the model projects high success rates for tasks spanning several hours, even under moderate

messiness. A more compact view of this progression is shown in Figure 3g, which visualizes the "rising tide" of automation for a fixed messiness level ($m = 4$). As compute and algorithmic efficiency grow, the maximum task length that can be reliably handled ($> 50\%$ success) moves from minutes to days.

Further, our Bayesian framework enables re-contextualization of traditional scaling laws into a framework grounded in human labor. Rather than plotting compute against abstract loss, we show the scaling relationship between training compute and task length (Figure 3h), across two performance regimes: reliable human assistance ($p = 0.5$) and near-full automation ($p = 0.99$), for tasks of varying complexity (messiness=$2, 4, 8$). This visualization reveals that increasing the length of a task or its messiness requires an exponential increase in compute to maintain the same success rate, potentially providing a map for not only model developers but also wider audience such as investors and policymakers to estimate the resource requirements necessary to automate specific sectors of the economy.

2.1.5 Key Differences to the METR Framework

To contextualize our approach, we provide a detailed comparison with the METR framework. While both methodologies aim to forecast AI task automation capabilities, they differ fundamentally in their modeling approaches and the information they utilize.

The METR framework fits individual logistic regression models for each AI system separately, predicting task success from task length alone. For each model, they extract the 50% task-completion time horizon (the task length at which success probability equals 50%) and fit an exponential trendline across models over time to project future capabilities (see Supplementary Information; Figure S.3).

In contrast, our Bayesian hierarchical framework jointly models all systems simultaneously, predicting task outcomes as a function of task length, task messiness, and effective training

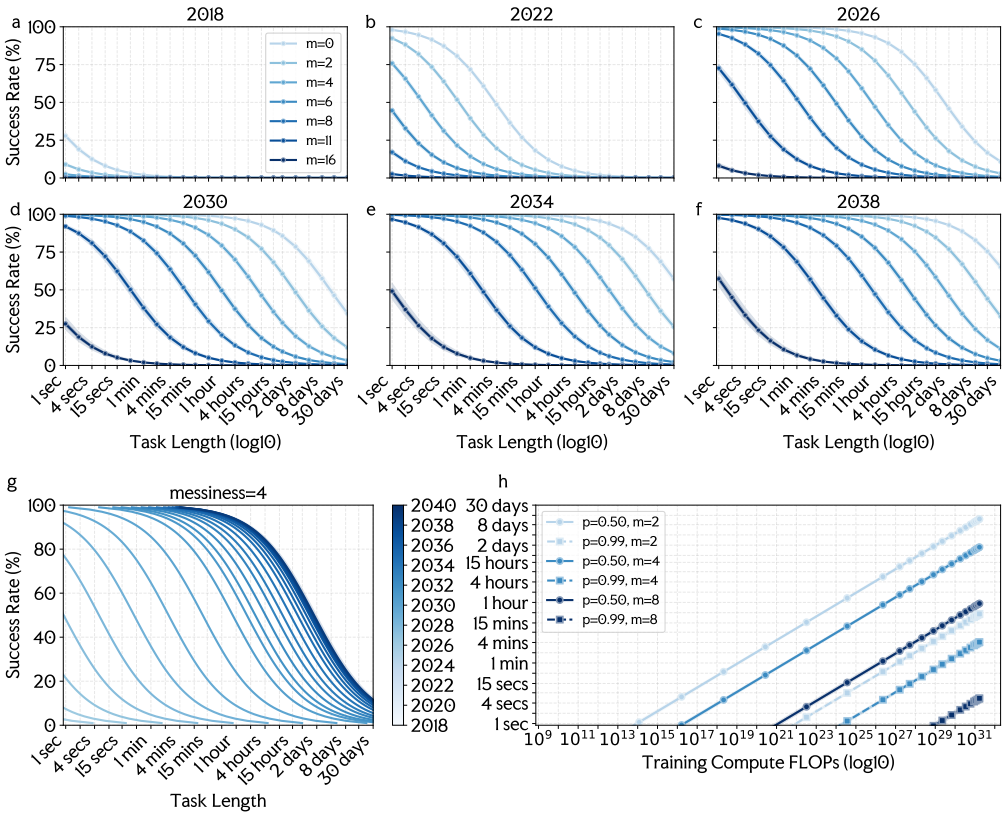

Figure 3: Projected automation success frontiers. (a–f) Evolution of the automation success frontier across time, showing task success rate as a function of task length and messiness. The frontier shifts rightward over time, with high success rates extending to longer and more complex tasks as effective compute grows. (g) The "rising tide" of automation for a fixed messiness level ($m = 4$), showing how the maximum task length achievable at $> 50\%$ success progresses from minutes to days between 2018 and 2040; the blue-shaded region highlights the constrained expansion of capabilities due to slowing infrastructure growth. (h) Scaling relationship between training compute and maximum automatable task length at two success thresholds ($p = 0.5$ and $p = 0.99$) across varying messiness levels ($m = 2, 4, 8$), revealing that maintaining a fixed success rate for longer or messier tasks requires exponentially more compute. Shaded bands represent 50% and 95% credible intervals throughout.

compute—the latter constrained by physical power availability, infrastructure efficiency, and algorithmic progress saturation. This integrated approach leverages previously disconnected datasets (power capacity projections, frontier model training histories, algorithmic efficiency benchmarks, and task performance data) to propagate uncertainty across the entire causal chain from infrastructure constraints to task-specific outcomes.

These contrasting approaches yield different forecast patterns subject to different constraint scenarios. Additionally, modeling messiness as a continuous variable enables us to generate separate projection profiles across task complexity levels (Figure 3), whereas METR's single trendline cannot capture such distinctions. However, direct head-to-head comparison requires caution: METR fits individual logistic regressions per model and derives a trendline from these separate estimates, while we jointly model all systems with shared constraints. For the closest comparison, Figure S.3 shows our forecast at average METR messiness (3.2/16). Beyond differences attributable to constraint scenarios, we also observe a discrepancy in trendline slopes prior to constraints becoming binding, which we attribute to the joint modeling objective pooling information across systems rather than fitting each independently. Our primary contribution builds on METR's foundational work by introducing a causally grounded approach that explicitly incorporates infrastructure constraints, algorithmic efficiency limits, and task complexity dimensions—providing a richer, more differentiated view of AI automation capabilities and their evolution over time, while enabling systematic scenario analysis where stakeholders can test alternative assumptions about each constraint.

2.2 Scenario Analysis

Having traced the components of our framework under a reference scenario, we further demonstrate its utility for scenario analysis, where different assumptions about scaling constraints can be entertained and forecasts compared. We consider three dimensions of uncertainty: algorith-

mic efficiency trajectories, the share of datacenter capacity invested in AI training, and total power supply growth.

First, we examine three algorithmic improvement scenarios in which the parametric form of efficiency gain follows a scaled sigmoid, a sum of shifted sigmoids, or an exponential function. As shown in Figure 4a–c, while each functional form similarly captures the historical trend of improvement when fitted to ECI (Epoch Capabilities Index; EpochAI 2025) scores (Table S.1), their future forecasts diverge substantially—with the exponential scenario projecting continued rapid gains while the sigmoid curves saturate. This divergence translates directly into meaningfully different automation time horizons (Figure 4d).

Second, we vary the share of datacenter capacity dedicated to AI training. In addition to the baseline share estimated from current data (1.74%, details see Supplementary Information), we consider two hypothetical cases in which the share grows linearly toward a target percentage over 20 years (3% and 8%). While these are stylized scenarios, they illustrate that increases in investment share can provide uplift to automation time horizons (Figure 4e–h).

Finally, we consider the effect of total power supply growth by substituting the EIA's baseline, high, and low economic growth trajectories (EIA, 2024). As shown in Figure 4i–k, different growth scenarios have minimal effect on automation time horizons. Under the baseline datacenter share, aggregate capacity growth appears insufficient to keep pace with the exponentially growing compute requirements of frontier AI training.

3 Discussion

In this manuscript, we proposed a novel framework of forecasting AI automation capabilities by explicitly modeling the causal relationship from fundamental infrastructure constraints—electrical power, efficiency improvements, and algorithmic progress—to task-specific performance outcomes. While previous approaches have treated these factors in isolation or relied on naive tem-

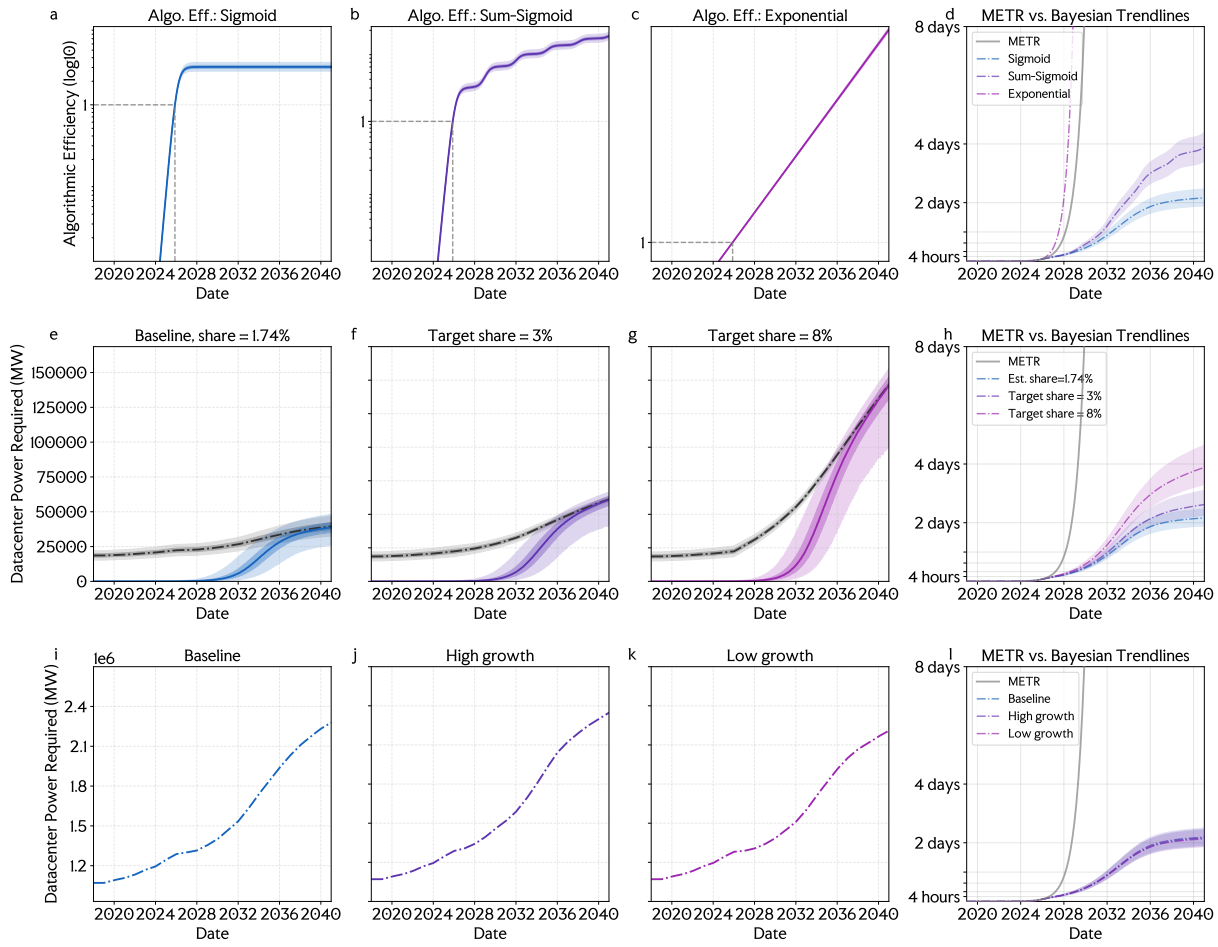

Figure 4: Scenario analysis across algorithmic efficiency, investment share, and power supply trajectories. (a–c) Three parametric forms of algorithmic efficiency—scaled sigmoid, sum of shifted sigmoids, and exponential—each fitted to historical ECI (Epoch Capabilities Index; EpochAI 2025) scores, with forecasts diverging substantially beyond the observation window. (d) Resulting automation time horizons under each algorithmic efficiency scenario. (e–g) Datacenter training power under three AI investment share scenarios: baseline (estimated current share), and two hypothetical cases growing linearly toward 3% and 8% of total datacenter capacity over 20 years. (h) Corresponding automation time horizons under each share scenario. (i–k) Datacenter training power under EIA baseline, high, and low economic growth trajectories for total U.S. power supply. (l) Corresponding automation time horizons, showing minimal divergence across growth scenarios. Shaded bands represent 50% and 95% credible intervals throughout.

poral extrapolations, our hierarchical Bayesian framework coherently propagates uncertainty across all components, enabling more grounded predictions that respect real-world boundaries.

The results demonstrate that future automation trajectories depend critically on how key constraints evolve. Our modular framework enables systematic scenario analysis across dimensions such as algorithmic efficiency trajectories, datacenter investment share, and power supply growth—each yielding meaningfully different time horizons. No single trajectory is inevitable; outcomes depend on the interplay of infrastructure, resource allocation, and algorithmic innovation. Crucially, the scenarios examined here are illustrative rather than exhaustive. The framework's design means that any component—whether a new estimate of power availability, a revised model of algorithmic progress, or an entirely new constraint—can be swapped in independently. As empirical understanding of AI development matures, the framework can be readily extended and adapted, providing a living tool for researchers, policymakers, and industry stakeholders to systematically interrogate how their own assumptions translate into automation timelines.

3.1 Modeling Choices and Opportunities for Extension

While our framework is modular and flexible—allowing uncertainty to propagate naturally through Bayesian inference—several design choices warrant careful consideration in light of emerging evidence. For example, we model algorithmic efficiency improvements as following three different parametric forms (i.e., sigmoid, exponential or sum-of-sigmoid). Alternative modeling choices for algorithmic improvement could be considered, such as using an economic model of ideas where algorithmic progress depends on experimental compute investment (Whitfill et al., 2025). Importantly, any functional form for algorithmic improvement imposes assumptions about future trajectories. Our framework's modularity addresses this limitation by treating algorithmic efficiency as an adjustable component: users can substitute alterna-

tive parameterizations or manually specify counterfactual scenarios to explore how different assumptions about algorithmic progress affect automation timelines.

In the current work, we omit training data as an explicit constraint. The role of data in limiting AI progress remains contested: while some analyses suggest that high-quality pretraining data may be approaching exhaustion (Villalobos et al., 2024), others argue that synthetic data generation, dedicated data collection efforts, and reinforcement learning in rich simulated environments provide effectively unlimited training signals—particularly as post-training has become increasingly central to frontier model development (Lambert, 2024; Kuba et al., 2025; Nadas et al., 2025; Niekerk et al., 2025). Existing forecasting work has often chosen to omit data constraints as a reasonable simplification (Erdil et al., 2025), and we followed this precedent. However, future extensions of our framework could incorporate data availability as an additional constraint, and our modular structure facilitates such additions.

3.2 Choice of Tasks, Beyond Software Engineering

Another limitation of our modeling approach stems from its reliance on METR's software engineering benchmarks for task automation modeling, which comes with two limitations. First, recent METR updates suggest that automated grading may inflate model success rates: approximately 24% of AI solutions that passed automated evaluation would have been rejected by human experts (Whitfill et al., 2026). If this finding generalises, it implies that AI performance is systematically overestimated, which would flatten the automation trendline. For the present contribution, we prioritize consistency with the original METR framework to best showcase the value added by our Bayesian approach, and therefore do not incorporate this alternative evaluation approach. Second, while software engineering tasks offer the advantage of relatively well-defined success criteria, making automation success tractable to model, these tasks may not fully represent other kinds of cognitive labor. Many economically valuable tasks across

other domains lack such clear criteria or are inherently open-ended.

Recent work has expanded to more diverse task sets as well as alternative task performance evaluation. For example, Patwardhan et al. (2025) proposed a benchmark spanning 44 occupations across the top nine sectors contributing to U.S. GDP, including finance, manufacturing, healthcare, and professional services. Rather than relying on binary success criteria, Patwardhan et al. (2025) employs expert-judged win rates comparing AI deliverables against work produced by human professionals with an average of 14 years of experience, accommodating the subjective and multifaceted nature of quality assessment in real-world work. Win rates may better capture automation's economic impact than binary success rates, as they can be more readily interpreted as a decision to substitute human work with work done by AI. Our Bayesian is designed to be flexible so that it can readily accommodate alternative task performance metrics as they emerge.

Similarly, Mazeika et al. (2025) propose the Remote Labor Index (RLI), comprising 240 end-to-end freelance projects spanning 23 categories of digital work, with completion times averaging 28.9 hours. Crucially, results show current frontier agents achieve automation rates below 2.5%, performing near the floor on this benchmark. This stark gap between performance on simplified computer-use evaluations and success on complete, economically valuable projects underscores that robust forecasting must grapple with the full complexity and diversity of real-world work.

A further limitation of current forecasting approaches—including our own—is the focus on fully autonomous AI performance. Before achieving complete automation, a more realistic scenario involves humans and AI working collaboratively. Recent empirical evidence presents a mixed picture: Becker et al. (2025) find that AI tools actually slowed experienced developers working on familiar codebases, while Patwardhan et al. (2025) demonstrate productivity gains when AI assistance is coupled with expert oversight. These contrasting findings point to the

complexity of measuring cooperative work.

A substantial body of work in organizational economics details many of these complexities, of which we highlight three: first, tasks are bundled into jobs and are governed within firms precisely because they cannot be fully specified or because there is economic benefit in retaining an adaptive job scope (Williamson, 1979; Simon, 1951). Second, the scope and assignment of these bundles are constrained meaningfully by communication and coordination costs (Becker and Murphy, 1992). Third, in joint production, the contribution of any individual agent (human or AI) to total output cannot be straightforwardly isolated from the contributions of others. (Holmström, 1982). Current forecasting and productivity research does not adequately model these dynamics. Future work should incorporate metrics capturing both the nature and intensity of human supervision and human-AI cooperation at different capability levels; our framework is well-positioned to accommodate these richer performance metrics as evaluation methodology matures.

3.3 Alternative Forecasting Paradigms

Our focus on aggregate task completion rates, while tractable to model, may obscure critical dynamics in how AI capabilities actually develop. Ord (2025) demonstrates that agent performance on METR's benchmark exhibits exponential decay with task duration, characterized by a "constant hazard rate"—agents fail at a relatively constant rate per unit of task time. This pattern implies that current systems primarily improve through reduced per-step error rates rather than developing capabilities like error detection or recovery (cf. Meyerson et al. 2025). This observation motivates an alternative forecasting paradigm: directly analyzing agent transcripts to track specific failure modes—goal persistence, information reconciliation, tool usage accuracy, and knowledge grounding (Luo, 2025). Importantly, failure mode analysis and resource-constraint modeling are complementary and integrating both could yield more comprehensive forecasts.

Finally, while our exponential algorithmic efficiency scenario can be interpreted as capturing a paradigm shift in AI development, our framework does not model more radical discontinuities. One emerging example is test-time scaling, where inference-time compute yields substantial capability gains largely independent of training-time resource constraints (Snell et al., 2024; OpenAI et al., 2024)—a dynamic our framework only partially captures through algorithmic efficiency improvements. More speculatively, recursive self-improvement—whereby AI systems autonomously enhance their own capabilities—could decouple automation trajectories from infrastructure constraints even more (Yuan et al., 2025; Simonds and Yoshiyama, 2025). Should such dynamics scale, they would represent a discontinuity that resource-based forecasting frameworks cannot anticipate.

3.4 Technical Feasibility, Adoption Lag and Economic Impact

Our results characterize the technical feasibility frontier for automating real-world software engineering tasks under three input constraints. However, technical feasibility is only one condition for real-world impact; others include adoption speed, the profitability of automating tasks, and which human skills are automated first.

3.4.1 Adoption Speed

For AI to matter economically, it must be adopted at scale to augment or replace work that produces economic value. Frontier labs have already made considerable efforts to classify AI use. Anthropic's Economic Index (Appel et al., 2026) measures how AI is being used across industries and geographies while OpenAI has produced work measuring both the occupation-level exposure estimates and, as already noted, task-level benchmarking of model performance on economically valuable work which explicitly bridges the gap between AI capability and economically relevant decision-making (Eloundou et al., 2023; Patwardhan et al., 2025). However,

enterprise adoption in the US remains low around 10 percent of enterprises as of September 2025, having grown by just 4 percentage points in the previous year (Bureau., 2025).

3.4.2 Productivity and Profitability of Automation

We also need to understand what the economic implications for a given level of capability and adoption would be. Existing work has already attempted to do so. At the broadest macroeconomic measure, Acemoglu (2025) estimates a 0.66 percent increase in Total Factor Productivity (TFP) which in turn maps to a mere 0.93 - 1.16 percent increase in GDP over a ten year period. On the more optimistic side, Goldman Sachs estimate GDP increase of 7 percent over a decade (Sachs, 2023). Acemoglu (2025)'s more conservative estimate assumes that just 4.6 percent of tasks are profitability exposed to AI (20 percent of tasks being exposed, of which only 23 percent are assumed to be profitably automatable) with an average cost savings of 14.4 percent. If the exponential increase in AI capability maps to exponential decreases to cost, clearly the potential savings could be much larger. We expect the inputs to these estimates have already changed significantly and believe a more sophisticated mapping of AI capability increases to cost reductions (in both breadth and depth) would be valuable.

3.4.3 Future Extensions of the Model

A natural extension to this study would be to show that the framework extends to tasks beyond software engineering into other economic domains (e.g., finance, healthcare, legal). This is important because the speed and shape of automation likely varies by domain due to different task requirements and constraints such as the distribution of messiness or cost of errors. For AI to be taken seriously as a transformative, general purpose technology, credible evidence of capability improvement on a wider set of economic domains is needed.

Our intention is to update the model over time on alternative technical assumptions and task domains (as in GDPval; Patwardhan et al. 2025). At this relatively early stage of AI

development, we believe a flexible forecasting model can help researchers, policymakers and capital allocators reason not only about whether tasks become automatable, but where and how fast transition pressure emerge across sectors.

4 Methods

4.1 Overview

The overall modeling logic for forecasting AI task automation is to conceptualize automation success as a function of both task-dependent variables (such as the length and complexity of the task) and time-dependent variables (such as infrastructure constraints, computational scaling, and algorithmic efficiency improvements). We developed a hierarchical Bayesian model to capture these causal relationships across multiple interconnected components (Figure 5).

4.2 Modeling Objective: Task Automation Success

We define the automation success of a task performed by a large language model (LLM) as conditioned on three key factors: (1) the time limit l for task completion (measured as the time needed by a human professional), (2) the messiness score m of the task, which serves as a proxy for task complexity as defined in prior work (Kwa et al., 2025), and (3) the training compute c estimated for the given LLM.

The probability p of successful task completion is modeled using a logistic regression framework:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot m + \beta_2 \cdot \log_{10}(l) + \beta_3 \cdot \log_{10}(c) \tag{1}$$

where the task outcome o (completed or not) follows a Bernoulli distribution:

$$o \sim \text{Bernoulli}(\text{logit}^{-1}(p)) \tag{2}$$

The parameters $\beta_0, \beta_1, \beta_2, \beta_3$ are given Normal priors with appropriate hyperparameters.

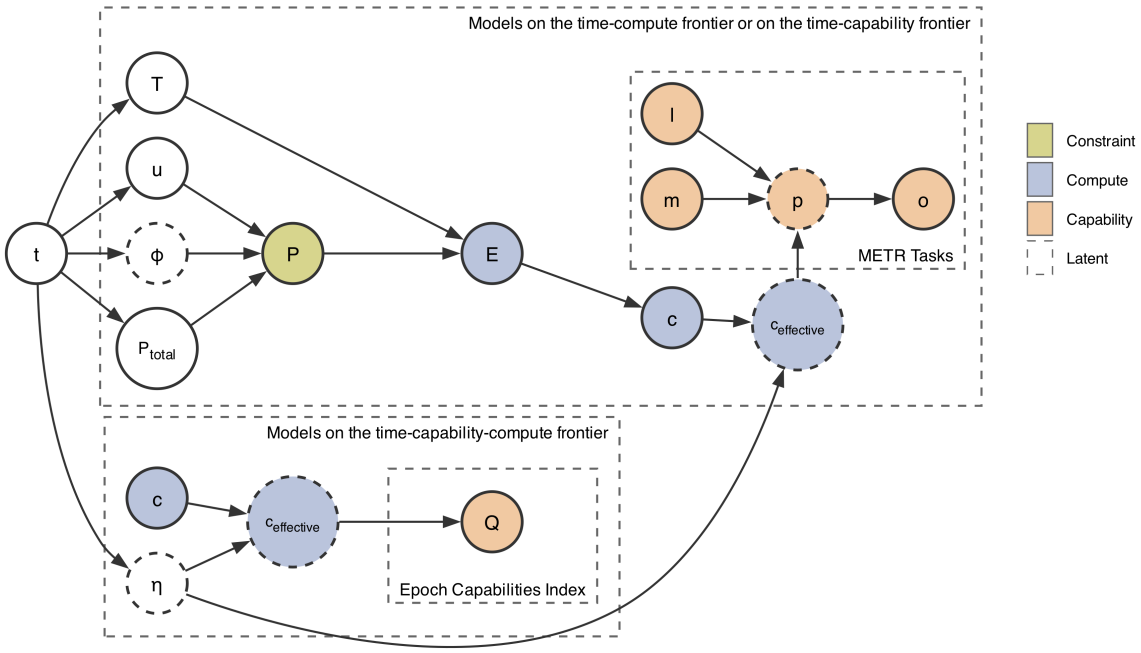

Figure 5: Causal structure of the hierarchical forecasting model. The framework consists of three interconnected layers. The **power-to-compute layer** (upper portion) models how time (t) and U.S. total power capacity (P_{total}) flow through allocation fractions (ϕ) and efficiency factor—power usage effectiveness (u)—to determine realized power (P), and through training duration (T) to determine training energy (E) and compute (c). The **algorithmic efficiency layer** (lower portion) converts raw compute (c) into effective compute ($c_{\text{effective}}$) via time-dependent algorithmic improvements (η), which relates to general capability (Q) measured by the Epoch Capabilities Index. The **task automation layer** (top-right) maps effective compute to task-specific outcomes (o) conditioned on task length (l) and messiness (m), producing success probabilities (p) for METR software engineering tasks. Dashed boxes indicate which model components are informed by different data partitions: models on the time-compute or time-capability frontier (upper box) inform power and compute scaling relationships, while models on the time-capability-compute frontier (lower box) inform algorithmic efficiency patterns. Node colors indicate variable types: constraints (green), compute measures (blue), capability measures (tan), and latent parameters (white with dashed borders). Arrows represent causal dependencies flowing from infrastructure constraints through computational resources to task outcomes.

4.3 Constraint Modeling: From Power to Compute

4.3.1 Power Availability and Allocation

A fundamental constraint on AI development is the availability of electrical power for training large-scale models. We model the power available for AI training as a fraction of total power capacity (P_{total}) in the United States, accounting for multiple allocation factors:

$$\text{share} = \frac{f_{\text{datacenter}} \cdot f_{\text{AI}} \cdot f_{\text{training}}}{n_{\text{hyperscalers}}} \tag{3}$$

$$P_{\text{bound}} = P_{\text{total}} \cdot \text{share} \tag{4}$$

where $f_{\text{datacenter}}$ represents the fraction of total power consumed by datacenters, f_{AI} is the fraction of datacenter power dedicated to AI workloads, f_{training} is the fraction of AI power used specifically for training (versus inference), and $n_{\text{hyperscalers}}$ accounts for the number of major companies competing for these resources. These fraction parameters are given Uniform priors encoding prior beliefs about power allocation derived from external knowledge about the industry. For details, see Supplementary Information.

4.3.2 Power Usage Effectiveness and Utilization

The actual power available for computation is modulated by two time-varying factors. First, the power usage effectiveness (PUE) $u(t_{\text{std}})$ decreases over time as datacenter infrastructure improves:

$$u(t_{\text{std}}) = u_{\text{floor}} + \exp(\alpha_0 + \alpha_1 \cdot t_{\text{std}}) \tag{5}$$

where t_{std} is standardized time, u_{floor} represents a lower bound on PUE (given a Uniform prior), and α_0 and α_1 are Normal-distributed parameters capturing the baseline and rate of improvement.

Second, the utilization fraction $\phi(t_{\text{std}})$ grows over time as more datacenter capacities are

allocated to training frontier models:

$$\phi(t_{\text{std}}) = \sigma(\delta_0 + \delta_1 \cdot t_{\text{std}}) \tag{6}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function. The parameters δ_0 and δ_1 (Normal and Log-Normal priors, respectively) control the baseline and growth rate of utilization.

4.3.3 Realized Power Draw

The realized power draw for AI training combines these factors:

$$\rho(t_{\text{std}}) = \frac{\phi(t_{\text{std}})}{u(t_{\text{std}})}, \quad P = \rho(t_{\text{std}}) \cdot P_{\text{bound}} \tag{7}$$

This parameterization ensures $P < P_{\text{bound}}$ by construction: since $\phi(t_{\text{std}}) \in (0, 1)$ (being a sigmoid) and $u(t_{\text{std}}) > 1$ (since $u(t_{\text{std}}) = u_{\text{floor}} + \exp(\cdot)$ with $u_{\text{floor}} > 1$), we have $\rho(t_{\text{std}}) < 1$, guaranteeing the power constraint is satisfied.

The observation model for ρ operates in logit space to ensure inferred values remain in the valid range $(0, 1)$:

$$\text{logit}(\rho_{\text{obs}}) \sim \mathcal{N}(\text{logit}(\rho_t), \sigma_\rho) \tag{8}$$

where observed power P_{obs} is first transformed to $\rho_{\text{obs}} = P_{\text{obs}}/P_{\text{bound}}$, then the likelihood is evaluated on $\text{logit}(\rho_{\text{obs}})$. This logit-space formulation allows symmetric Gaussian noise in the unconstrained space while maintaining the proper bounds.

4.3.4 Training Duration

Training duration is modeled as a fraction of an upper bound that evolves sigmoidally over time:

$$\text{training\_frac}(t_{\text{std}}) = \sigma(\tau_0 + \tau_1 \cdot t_{\text{std}}) \tag{9}$$

$$T_{\text{days}} = T_{\text{upper}} \cdot \text{training\_frac}(t_{\text{std}}) \tag{10}$$

where T_{upper} represents the maximum plausible training duration and $\sigma(\cdot)$ denotes the logistic sigmoid function. The parameters τ_0 and τ_1 (with Normal priors) control the baseline and growth rate of training duration over time.

The observation model operates in log space:

$$\log_{10}(T_{\text{obs}}) \sim \mathcal{N}(\log_{10}(T_{\text{pred}}), \sigma_{\log T}) \tag{11}$$

where T_{upper} receives a Normal prior and $\sigma_{\log T}$ has a HalfNormal prior.

4.3.5 Energy-to-Compute Relationship

The training compute c is derived from energy expenditure, which combines power draw with training duration:

$$E = P \cdot T_{\text{days}} \tag{12}$$

where E represents energy in watt-days. The relationship from energy to training compute follows a log-linear form:

$$\log_{10}(c) = \gamma_0 + \gamma_1 \cdot \log_{10}(E) \tag{13}$$

with observation model:

$$\log_{10}(c_{\text{obs}}) \sim \mathcal{N}(\log_{10}(c_{\text{pred}}), \sigma_c) \tag{14}$$

The parameters γ_0 and γ_1 are given Normal priors, while σ_c (the observation noise) receives a HalfNormal prior.

4.4 Algorithmic Efficiency and Effective Compute

4.4.1 Efficiency Improvements Over Time

Beyond raw computational resources, algorithmic improvements increase the effective compute available for training. We model algorithmic efficiency $\eta(t_{\text{std}})$ as a function of standardized

time, normalized so that $\eta(t_{\text{current}}) = 1$, where t_{current} is the latest model release date (standardized) in the Epoch Capabilities Index dataset (EpochAI, 2025). We implement three parametric forms: a scaled sigmoid (bounded, saturating), an exponential (unbounded growth), and a sum of shifted sigmoids (bounded staircase-shaped improvement). The same capability and task-outcome likelihood is used for all forms; extrapolations beyond the observation period are driven by the choice of functional form and its priors.

Scaled sigmoid The default form is a scaled sigmoid:

$$\eta(t_{\text{std}}) = A \cdot \sigma(\eta_0 + \eta_1 \cdot t_{\text{std}}) \tag{15}$$

where η_0 and η_1 receive Normal and LogNormal priors, respectively, and the scaling factor A is determined by the constraint $\eta(t_{\text{current}}) = 1$:

$$A = \frac{1}{\sigma(\eta_0 + \eta_1 \cdot t_{\text{current}})} \tag{16}$$

Thus $\eta(t_{\text{std}}) < 1$ for $t_{\text{std}} < t_{\text{current}}$ (past models were less efficient) and $\eta(t_{\text{std}}) > 1$ for $t_{\text{std}} > t_{\text{current}}$, with an upper bound set by A. This formulation captures gradual accumulation of algorithmic gains with eventual saturation, and reflects uncertainty about whether current paradigms are near fundamental limits or allow substantial further gains.

Exponential An alternative is constant proportional growth in efficiency:

$$\eta(t_{\text{std}}) = \exp\big(\lambda \cdot (t_{\text{std}} - t_{\text{current}})\big) \tag{17}$$

where λ is the log-scale growth rate and receives a LogNormal prior. The constraint $\eta(t_{\text{current}}) = 1$ holds by construction. Past efficiency is $\eta(t_{\text{std}}) < 1$ for $t_{\text{std}} < t_{\text{current}}$; for future times $\eta(t_{\text{std}})$ grows without bound. This form is appropriate if one does not assume near-term saturation, at the cost of allowing large extrapolated efficiency unless the prior on λ is tightened.

Sum of shifted sigmoids To model episodic breakthroughs interspersed with periods of slower progress, we also consider a sum of K temporally shifted sigmoids sharing a common slope:

$$\eta(t_{\text{std}}) = A \cdot \sum_{k=0}^{K-1} \sigma\big(\eta_0 - k\,\eta_1\,\Delta + \eta_1 \cdot t_{\text{std}}\big) \tag{18}$$

where η_0 and η_1 receive Normal and LogNormal priors as in the scaled sigmoid, K is the fixed number of breakthroughs, $\Delta > 0$ is the fixed inter-breakthrough gap (in standardized time units), and the scaling factor A satisfies $\eta(t_{\text{current}}) = 1$:

$$A = \frac{1}{\displaystyle\sum_{k=0}^{K-1} \sigma\big(\eta_0 - k\,\eta_1\,\Delta + \eta_1 \cdot t_{\text{current}}\big)} \tag{19}$$

The $k = 0$ sigmoid is identical to the scaled sigmoid of Equation (15); each subsequent sigmoid is shifted later by Δ, placing the k-th midpoint at $m_k = -\eta_0/\eta_1 + k\,\Delta$. The first midpoint m_0 is thus data-informed through (η_0, η_1), while the positions of future breakthroughs are determined by the prior specification of Δ. The long-run ceiling is $A{\cdot}K$; for breakthroughs whose midpoints lie well in the future at t_{current}, this approaches K times the ceiling of the single sigmoid. The resulting staircase-shaped trajectory captures the intuition that algorithmic progress occurs in waves rather than continuously, with K and Δ acting as scenario parameters expressing beliefs about the number and pace of future breakthroughs rather than quantities the historical data can directly constrain.

4.4.2 Effective Compute

The effective compute combines raw compute with algorithmic efficiency:

$$c_{\text{effective}} = c \cdot \eta(t_{\text{std}}) \tag{20}$$

or equivalently in log space:

$$\log_{10}(c_{\text{effective}}) = \log_{10}(c) + \log_{10}(\eta) \tag{21}$$

4.4.3 Capability Score Model

We relate effective compute to model capability using a linear relationship in log-effective-compute space:

$$Q = \kappa_0 + \kappa_1 \cdot \log_{10}(c_{\text{effective}}) \tag{22}$$

with observation model:

$$Q_{\text{obs}} \sim \mathcal{N}(Q_{\text{pred}}, \sigma_{\text{cap}}) \tag{23}$$

where Q represents capability as measured by the Epoch Capabilities Index (EpochAI, 2025). The parameters κ_0 and κ_1 receive Normal priors, while σ_{cap} has a HalfNormal prior.

4.5 Model Selection and Data Partitioning

To ensure our model captures both the frontier of computational scaling and algorithmic efficiency, we partition the training data into two complementary sets:

4.5.1 Time-Compute or Time-Capability Frontier

Components related to power constraints and compute scaling are informed by models selected based on the following criteria: Models are included in this frontier if they satisfy either of two criteria: (1) they were among the top 5 models by training compute at their release date using the "top-N at release" method from Epoch/EPRI analysis, where models are ranked by the difference between their total training compute (including fine-tuning) and their base model's training compute (You et al., 2025) or (2) they lie on the 2D Pareto frontier minimizing release time while maximizing capability score on METR task automation evaluations (Kwa et al., 2025). This selection identifies models that represented the frontier of compute scale or task automation capability at the time of their release.

4.5.2 Time-Capability-Compute Frontier

Components related to algorithmic efficiency are informed by models that are selected by computing a 3D Pareto frontier that simultaneously minimizes training compute, minimizes release date, and maximizes capability score on the Epoch Capabilities Index (ECI; EpochAI 2025; Ho et al. 2025), which combines scores from many different AI benchmarks into a single general capability scale. This identifies models that achieved high capability with minimal compute—representing efficient use of training resources through algorithmic or architectural innovations rather than (just) brute-force scaling.

4.5.3 Summary of Observation Models

The full model combines multiple observation equations that link the latent generative model to observed data. Table 1 summarizes all observation likelihoods used in the model.

Table 1: Summary of observation models (likelihoods) linking latent model predictions to observed data.

Observation	**Likelihood**
ρ_{obs}	$\text{logit}(\rho_{\text{obs}}) \sim \mathcal{N}(\text{logit}(\rho_t), \sigma_\rho)$
c_{obs}	$\log_{10}(c_{\text{obs}}) \sim \mathcal{N}(\log_{10}(c_{\text{pred}}), \sigma_c)$
T_{obs}	$\log_{10}(T_{\text{obs}}) \sim \mathcal{N}(\log_{10}(T_{\text{pred}}), \sigma_{\log T})$
o	$o \sim \text{Bernoulli}(\text{logit}^{-1}(p))$
u_{obs} (PUE)	$u_{\text{obs}} \sim \mathcal{N}(u(t_{\text{std}}), \sigma_u)$
Q_{obs} (Capability)	$Q_{\text{obs}} \sim \mathcal{N}(Q_{\text{pred}}, \sigma_{\text{cap}})$

4.6 Data Sources

4.6.1 METR Tasks

The data we used to model the task automation success captured by Eq. 1 as part of the time-compute-capacity frontier comes from Kwa et al. (2025), which presents a benchmark dataset consisting of 170 tasks across three suites: HCAST (97 diverse software/cybersecurity/ML

tasks, 1 minute to 30 hours; Rein et al. 2025), RE-Bench (7 ML research engineering tasks, 8 hours each; Wijk et al. 2025), and Software Atomic Actions (SWAA; 66 single-step software tasks, < 1 minute). The authors evaluated 13 frontier models from GPT-2 (2019) to Claude 3.7 Sonnet (2025) on these tasks.

Task length was measured using human baseline completion times from skilled professionals (average 5 years experience), calculated as the geometric mean of successful attempts. Over 800 human baselines totaling 2,529 hours were collected, with 148 of 170 tasks having actual human baselines and 21 tasks using manual estimates. Task success was measured through automatic scoring functions, with most tasks scored binary (0 or 1). Continuously-scored tasks were binarized using task-specific thresholds representing human performance levels. Models performed 8 runs per task, with success rates averaged across runs.

Messiness was quantified using 16 binary factors capturing systematic differences from real-world tasks, including whether tasks were sourced from real contexts, involved resource constraints, dynamic environments, required real-time coordination, or had implicit requirements not explicitly stated. Tasks were scored 0-16 based on these factors, with mean messiness of 3.2/16 for HCAST and RE-Bench tasks.

For complete methodological details, including baseliner selection criteria, scaffolding implementations, and scoring procedures, please refer to the original paper (Kwa et al., 2025).

4.7 Power and Compute Estimates

4.7.1 Epoch AI Databases

The data we used to model power and compute relationships captured by Eq. 13-14 as part of the time-compute-capability frontier is obtained from two primary datasets from Epoch AI's analysis of frontier AI training power demands: the ML Systems Electricity dataset and the All AI Models database.

Following Epoch AI's methodology, we applied their power calibration approach based on NVIDIA DGX H100 system specifications to convert reported TDP measurements into training power draw estimates in watts. Training time and compute values were used to identify frontier models—defined as systems in the top-5 by training compute at their release date, following Epoch AI's frontier classification methodology for models released from 2018 onwards. See Epoch AI's analysis for full methodological details (You et al., 2025).

4.8 US Electric Power Capacity Data

We obtained data for total power capacity in the US from the Annual Energy Outlook published by the Energy Information Administration (EIA). Specifically, we used the **Electric power sector capacity** under the reference scenario, which includes plants that only produce electricity and combined heat and power plants that have a regulatory status.

The capacity measurements represent net summer capacity—the maximum output that generating equipment can supply to system load, adjusted for ambient summer conditions. We used EIA's projections based on a business-as-usual trend estimate, given known technology, technological, and demographic trends spanning 2019-2050 (EIA, 2024).

4.9 Epoch Capabilities Index

For modeling algorithmic efficiency captured by Eq. 15-23 (Sec. 4.4), we used data from Epoch Capabilities Index (EpochAI, 2025) where we obtained training compute in floating-point operations (FLOP) and ECI (Epoch Capabilities Index) scores—composite capability measurements.

To integrate capability scores with the broader model dataset above, we matched Epoch Capabilities Index entries to the All AI Models database using a fuzzy matching algorithm. When multiple matches existed, we prioritized the highest match score, followed by the highest

capability score. To ensure one-to-one mappings, we retained only the best capability entry per model in the All AI Models database.

Using the matched capability data, we identified models on 3D frontier as described above. See Epoch AI's Capabilities Index documentation for full details on ECI score construction (EpochAI, 2025).

4.10 Bayesian Inference

The full Bayesian model was implemented in NumPyro using a two-stage inference procedure. First, we employed stochastic variational inference (SVI) to optimize the hyperparameters of the prior distributions by maximizing the marginal likelihood of the observed data. This SVI stage learns appropriate centering and scaling for the priors of model parameters that govern functional relationships (e.g., power-to-compute scaling, algorithmic efficiency growth, capability-compute relationships). Critically, priors informed by external domain knowledge—specifically the power allocation fractions $f_{\text{datacenter}}$, f_{AI}, f_{training}, and $n_{\text{hyperscalers}}$, as well as the PUE floor u_{floor}—remain fixed during this optimization stage, as they encode real-world constraints rather than relationships to be learned from the data. Following hyperparameter optimization via SVI, we performed full Bayesian inference using Markov Chain Monte Carlo (MCMC) with the No-U-Turn Sampler (NUTS; Homan and Gelman 2014) to obtain posterior distributions for all model parameters.

Author Contributions

XL drafted the initial manuscript. HS and XL developed and implemented the Bayesian hierarchical framework. RA and SB performed exploratory pilot work that shaped the research approach. JM provided economic analysis and broader impact perspectives. MW initiated and defined the scope of the research project. All authors contributed to manuscript revision and

framework refinement.

Declaration of Interests

The authors declare no competing interests.

References

Acemoglu D (2025) The simple macroeconomics of AI. Economic Policy 40(121):13–58

Amazon (2025) How Amazon is helping to build one of the first modular nuclear reactor facilities in the United States. URL https://www.aboutamazon.com/news/sustainability/amazon-smr-nuclear-energy

Appel R, Massenkoff M, McCrory P, McCain M, Heller R, Neylon T, Tamkin A (2026) Anthropic Economic Index report: economic primitives. URL https://www.anthropic.com/research/anthropic-economic-index-january-2026-report

Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286(5439):509–512, DOI 10.1126/science.286.5439.509, URL https://www.science.org/doi/abs/10.1126/science.286.5439.509, _eprint: https://www.science.org/doi/pdf/10.1126/science.286.5439.509

Becker GS, Murphy KM (1992) The division of labor, coordination costs, and knowledge. The Quarterly Journal of Economics 107(4):1137–1160

Becker J, Rush N, Barnes E, Rein D (2025) Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. URL https://arxiv.org/abs/2507.09089, _eprint: 2507.09089

Bureau UC (2025) Business Trends and Outlook Survey: AI Core Questions [Data set]. URL https://www.census.gov/hfp/btos

Constellation (2024) Constellation to Launch Crane Clean Energy Center, Restoring Jobs and Carbon-Free Power to The Grid. URL https://www.constellationenergy.com/newsroom/2024/Constellation-to-Launch-Crane-Clean-Energy-Center-Restoring-Jobs-and-Car
html

Edelman Y, Ho A (2025) Compute scaling will slow down due to increasing lead times. URL https://epoch.ai/gradient-updates/compute-scaling-will-slow-down-due-to-increasing-lead-times

EIA (2024) Annual Energy Outlook 2025 Table: Table 9. Electricity Generating Capacity. URL https://www.eia.gov/outlooks/aeo/data/browser/#/?id=9-AEO2025&region=0-0&cases=ref2025&start=2023&end=2025&f=A&linechart=ref2025-d032025a.4-9-AEO2025&map=&sourcekey=0

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. DOI 10.48550/arXiv.2303.10130, URL http://arxiv.org/abs/2303.10130, arXiv:2303.10130 [econ]

EpochAI (2025) Epoch Capabilities Index. URL https://epoch.ai/benchmarks/eci

Erdil E (2024) Optimally allocating compute between inference and training. URL https://epoch.ai/blog/optimally-allocating-compute-between-inference-and-training

Erdil E, Potlogea A, Besiroglu T, Roldan E, Ho A, Sevilla J, Barnett M, Vrzla M, Sandler R

(2025) GATE: An Integrated Assessment Model for AI Automation. DOI 10.48550/arXiv.2503.04941, URL http://arxiv.org/abs/2503.04941, arXiv:2503.04941 [econ]

Geman B (2025) Ontario modular reactor to be first in "Western world," GE predicts. URL https://www.axios.com/2025/05/08/ontario-small-modular-reactor-to-be-first

Ho A, Denain JS, Atanasov D, Albanie S, Shah R (2025) A Rosetta Stone for AI Benchmarks. DOI 10.48550/arXiv.2512.00193, URL http://arxiv.org/abs/2512.00193, arXiv:2512.00193 [cs]

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Casas DdL, Hendricks LA, Welbl J, Clark A, Hennigan T, Noland E, Millican K, Driessche Gvd, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Rae JW, Vinyals O, Sifre L (2022) Training Compute-Optimal Large Language Models. DOI 10.48550/arXiv.2203.15556, URL http://arxiv.org/abs/2203.15556, arXiv:2203.15556 [cs]

Holmström B (1982) Moral hazard in teams. The Bell Journal of Economics 13(2):324–340, DOI 10.2307/3003457

Homan MD, Gelman A (2014) The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res 15(1):1593–1623, publisher: JMLR.org

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling Laws for Neural Language Models. DOI 10.48550/arXiv.2001.08361, URL http://arxiv.org/abs/2001.08361, arXiv:2001.08361 [cs]

Kuba JG, Gu M, Ma Q, Tian Y, Mohan V, Chen J (2025) Language Self-Play For Data-Free Training. DOI 10.48550/arXiv.2509.07414, URL http://arxiv.org/abs/2509.07414, arXiv:2509.07414 [cs]

Kwa T, West B, Becker J, Deng A, Garcia K, Hasin M, Jawhar S, Kinniment M, Rush N, Arx SV, Bloom R, Broadley T, Du H, Goodrich B, Jurkovic N, Miles LH, Nix S, Lin T, Parikh N, Rein D, Sato LJK, Wijk H, Ziegler DM, Barnes E, Chan L (2025) Measuring AI Ability to Complete Long Tasks. DOI 10.48550/arXiv.2503.14499, URL http://arxiv.org/abs/2503.14499, arXiv:2503.14499 [cs]

Lambert N (2024) A recipe for frontier model post-training. URL https://www.interconnects.ai/p/frontier-model-post-training

Leppert R (2025) What we know about energy use at U.S. data centers amid the AI boom. URL https://www.pewresearch.org/short-reads/2025/10/24/what-we-know-about-energy-use-at-us-data-centers-amid-the-ai-boom/

Luo X (2025) Towards Tractable and High-Resolution Agent Capability Forecasting: A Research Proposal. DOI 10.13140/RG.2.2.32954.04801, URL https://www.researchgate.net/doi/10.13140/RG.2.2.32954.04801

Mazeika M, Gatti A, Menghini C, Sehwag UM, Singhal S, Orlovskiy Y, Basart S, Sharma M, Peskoff D, Lau E, Lim J, Carroll L, Blair A, Sivakumar V, Basu S, Kenstler B, Ma Y, Michael J, Li X, Ingebretsen O, Mehta A, Mottola J, Teichmann J, Yu K, Shaik Z, Khoja A, Ren R, Hausenloy J, Phan L, Htet Y, Aich A, Rabbani T, Shah V, Novykov A, Binder F, Chugunov K, Ramirez L, Geralnik M, Mesura H, Lee D, Cardona EYH, Diamond A, Yue S, Wang A, Liu B, Hernandez E, Hendrycks D (2025) Remote Labor Index: Measuring AI Automation of Remote Work. DOI 10.48550/arXiv.2510.26787, URL http://arxiv.org/abs/2510.26787, arXiv:2510.26787 [cs]

Meyerson E, Paolo G, Dailey R, Shahrzad H, Francon O, Hayes CF, Qiu X, Hodjat B, Miikku-

lainen R (2025) Solving a Million-Step LLM Task with Zero Errors. DOI 10.48550/arXiv.2511.09030, URL http://arxiv.org/abs/2511.09030, arXiv:2511.09030 [cs]

Moss S (2024) OpenAI training and inference costs could reach $7bn for 2024, AI startup set to lose $5bn - report. URL https://www.datacenterdynamics.com/en/news/openai-training-and-inference-costs-could-reach-7bn-for-2024-ai-startup-

Nadas M, Diosan L, Tomescu A (2025) Synthetic Data Generation Using Large Language Models: Advances in Text and Code. IEEE Access 13:134615–134633, DOI 10.1109/ACCESS.2025.3589503, URL http://arxiv.org/abs/2503.14023, arXiv:2503.14023 [cs]

NEA (2025) NEA Small Modular Reactor (SMR) Dashboard. URL https://www.oecd-nea.org/jcms/pl_73678/nea-small-modular-reactor-smr-dashboard

Niekerk Cv, Vukovic R, Ruppik BM, Lin Hc, Gašić M (2025) Post-Training Large Language Models via Reinforcement Learning from Self-Feedback. DOI 10.48550/arXiv.2507.21931, URL http://arxiv.org/abs/2507.21931, arXiv:2507.21931 [cs]

OpenAI, :, Jaech A, Kalai A, Lerer A, Richardson A, El-Kishky A, Low A, Helyar A, Madry A, Beutel A, Carney A, Iftimie A, Karpenko A, Passos AT, Neitz A, Prokofiev A, Wei A, Tam A, Bennett A, Kumar A, Saraiva A, Vallone A, Duberstein A, Kondrich A, Mishchenko A, Applebaum A, Jiang A, Nair A, Zoph B, Ghorbani B, Rossen B, Sokolowsky B, Barak B, McGrew B, Minaiev B, Hao B, Baker B, Houghton B, McKinzie B, Eastman B, Lugaresi C, Bassin C, Hudson C, Li CM, Bourcy Cd, Voss C, Shen C, Zhang C, Koch C, Orsinger C, Hesse C, Fischer C, Chan C, Roberts D, Kappler D, Levy D, Selsam D, Dohan D, Farhi D, Mely D, Robinson D, Tsipras D, Li D, Oprica D, Freeman E, Zhang E, Wong E, Proehl E, Cheung E, Mitchell E, Wallace E, Ritter E, Mays E, Wang F, Such FP, Raso F, Leoni F,

Tsimpourlas F, Song F, Lohmann Fv, Sulit F, Salmon G, Parascandolo G, Chabot G, Zhao G, Brockman G, Leclerc G, Salman H, Bao H, Sheng H, Andrin H, Bagherinezhad H, Ren H, Lightman H, Chung HW, Kivlichan I, O'Connell I, Osband I, Gilaberte IC, Akkaya I, Kostrikov I, Sutskever I, Kofman I, Pachocki J, Lennon J, Wei J, Harb J, Twore J, Feng J, Yu J, Weng J, Tang J, Yu J, Candela JQ, Palermo J, Parish J, Heidecke J, Hallman J, Rizzo J, Gordon J, Uesato J, Ward J, Huizinga J, Wang J, Chen K, Xiao K, Singhal K, Nguyen K, Cobbe K, Shi K, Wood K, Rimbach K, Gu-Lemberg K, Liu K, Lu K, Stone K, Yu K, Ahmad L, Yang L, Liu L, Maksin L, Ho L, Fedus L, Weng L, Li L, McCallum L, Held L, Kuhn L, Kondraciuk L, Kaiser L, Metz L, Boyd M, Trebacz M, Joglekar M, Chen M, Tintor M, Meyer M, Jones M, Kaufer M, Schwarzer M, Shah M, Yatbaz M, Guan MY, Xu M, Yan M, Glaese M, Chen M, Lampe M, Malek M, Wang M, Fradin M, McClay M, Pavlov M, Wang M, Wang M, Murati M, Bavarian M, Rohaninejad M, McAleese N, Chowdhury N, Chowdhury N, Ryder N, Tezak N, Brown N, Nachum O, Boiko O, Murk O, Watkins O, Chao P, Ashbourne P, Izmailov P, Zhokhov P, Dias R, Arora R, Lin R, Lopes RG, Gaon R, Miyara R, Leike R, Hwang R, Garg R, Brown R, James R, Shu R, Cheu R, Greene R, Jain S, Altman S, Toizer S, Toyer S, Miserendino S, Agarwal S, Hernandez S, Baker S, McKinney S, Yan S, Zhao S, Hu S, Santurkar S, Chaudhuri SR, Zhang S, Fu S, Papay S, Lin S, Balaji S, Sanjeev S, Sidor S, Broda T, Clark A, Wang T, Gordon T, Sanders T, Patwardhan T, Sottiaux T, Degry T, Dimson T, Zheng T, Garipov T, Stasi T, Bansal T, Creech T, Peterson T, Eloundou T, Qi V, Kosaraju V, Monaco V, Pong V, Fomenko V, Zheng W, Zhou W, McCabe W, Zaremba W, Dubois Y, Lu Y, Chen Y, Cha Y, Bai Y, He Y, Zhang Y, Wang Y, Shao Z, Li Z (2024) OpenAI o1 System Card. URL https://arxiv.org/abs/2412.16720, _eprint: 2412.16720

Ord T (2025) https://www.tobyord.com/writing/half-life. URL https://www.tobyord.com/writing/half-life

Patwardhan T, Dias R, Proehl E, Kim G, Wang M, Watkins O, Fishman SP, Aljubeh M, Thacker P, Fauconnet L, Kim NS, Chao P, Miserendino S, Chabot G, Li D, Sharman M, Barr A, Glaese A, Tworek J (2025) GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. DOI 10.48550/arXiv.2510.04374, URL http://arxiv.org/abs/2510.04374, arXiv:2510.04374 [cs]

Rein D, Becker J, Deng A, Nix S, Canal C, O'Connel D, Arnott P, Bloom R, Broadley T, Garcia K, Goodrich B, Hasin M, Jawhar S, Kinniment M, Kwa T, Lajko A, Rush N, Sato LJK, Arx SV, West B, Chan L, Barnes E (2025) HCAST: Human-Calibrated Autonomy Software Tasks. URL https://arxiv.org/abs/2503.17354, _eprint: 2503.17354

Sachs G (2023) Generative AI could raise global GDP by 7 percent. URL https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

Simon HA (1951) A formal theory of the employment relationship. Econometrica 19(3):293–305

Simonds T, Yoshiyama A (2025) LADDER: Self-Improving LLMs Through Recursive Problem Decomposition. DOI 10.48550/arXiv.2503.00735, URL http://arxiv.org/abs/2503.00735, arXiv:2503.00735 [cs]

Snell C, Lee J, Xu K, Kumar A (2024) Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. DOI 10.48550/arXiv.2408.03314, URL http://arxiv.org/abs/2408.03314, arXiv:2408.03314 [cs]

Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M (2024) Will we run out of data? Limits of LLM scaling based on human-generated data. DOI 10.48550/arXiv.2211.04325, URL http://arxiv.org/abs/2211.04325, arXiv:2211.04325 [cs]

Vries-Gao Ad (2025) Artificial intelligence: Supply chain constraints and energy implications. Joule 9(6):101961, DOI https://doi.org/10.1016/j.joule.2025.101961, URL https://www.sciencedirect.com/science/article/pii/S2542435125001424

West GB, Brown JH, Enquist BJ (1997) A General Model for the Origin of Allometric Scaling Laws in Biology. Science 276(5309):122–126, DOI 10.1126/science.276.5309.122, URL https://www.science.org/doi/abs/10.1126/science.276.5309.122, _eprint: https://www.science.org/doi/pdf/10.1126/science.276.5309.122

Whitfill P, Snodin B, Becker J (2025) Forecasting AI Time Horizon Under Compute Slowdowns. DOI 10.48550/arXiv.2511.19492, URL http://arxiv.org/abs/2511.19492, arXiv:2511.19492 [cs]

Whitfill P, Wu C, Becker J, Rush N (2026) Many SWE-bench-Passing PRs Would Not Be Merged into Main. URL https://metr.org/notes/2026-03-10-many-swe-bench-passing-prs-would-not-be-merged-into-main/

Wijk H, Lin T, Becker J, Jawhar S, Parikh N, Broadley T, Chan L, Chen M, Clymer J, Dhyani J, Ericheva E, Garcia K, Goodrich B, Jurkovic N, Karnofsky H, Kinniment M, Lajko A, Nix S, Sato L, Saunders W, Taran M, West B, Barnes E (2025) RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. URL https://arxiv.org/abs/2411.15114, _eprint: 2411.15114

Williamson OE (1979) Transaction-cost economics: The governance of contractual relations. Journal of Law and Economics 22(2):233–261

You J, Owen D, Porter D, Wilson T (2025) Scaling Intelligence: The Exponential Growth of AI's Power Needs. URL https://www.epri.com/research/products/000000003002033669, issue: 3002033669

Yuan W, Pang RY, Cho K, Li X, Sukhbaatar S, Xu J, Weston J (2025) Self-Rewarding Language Models. DOI 10.48550/arXiv.2401.10020, URL http://arxiv.org/abs/2401.10020, arXiv:2401.10020 [cs]

Supplementary Information

Power Fraction Estimates

We have historical datacenter power draws used to train frontier models and EIA estimates for total U.S. power capacity with forecasts through 2050. We model the power bound P_{bound}, which represents the maximal power at a given date that can be dedicated to training a single frontier model. Following the constraint modeling framework, the power bound is calculated as:

$$\text{share} = \frac{f_{\text{datacenter}} \cdot f_{\text{AI}} \cdot f_{\text{training}}}{n_{\text{hyperscalers}}} \tag{24}$$

$$P_{\text{bound}} = P_{\text{total}} \cdot \text{share} \tag{25}$$

Where:

- P_{total} = total U.S. power capacity
- $f_{\text{datacenter}}$ = fraction of total power consumed by datacenters
- f_{AI} = fraction of datacenter power dedicated to AI workloads
- f_{training} = fraction of AI power used specifically for training (versus inference)
- $n_{\text{hyperscalers}}$ = number of major companies competing for these resources

Estimating $f_{\text{datacenter}}$

Leppert (2025) gives the total energy consumption by U.S. data centers both historically and projected into 2030. It also states that 2024 usage was "more than 4%" of total US power usage. Using this information, we can get the following table.

Year	$f_{\text{datacenter}}$ (%)
2020	2.36
2023	3.67
2024	4.00
2030	9.31

Fitting an exponential to this data gives a roughly 14% annual growth rate in datacenter electricity usage, with a further projection into 2040 giving a total fraction of 37%. If we want to estimate a single fraction to act for the purposes of the envelope, a value between 10% and 37% seems feasible.

Assumptions and limitations

- U.S. power capacity is also growing, so the fraction used by datacenters given above could be overestimates. However a constant assumption on total power capacity seems reasonable if power capacity is growing much more slowly than datacenter usage.
- There is unprecedented build-out of infrastructure specifically for datacenters, including dedicated power generation. Major tech companies are investing in nuclear energy: Microsoft has signed a 20-year agreement to restart the Three Mile Island Unit 1 reactor (expected operational by 2028; Constellation 2024, and Amazon is co-developing a modular nuclear facility using X-energy SMRs in Washington state (Amazon, 2025). However, small modular reactors (SMRs) have long deployment timelines—the first commercial SMR in the Western world is projected for Ontario by 2030 (Geman, 2025), and U.S. commercial SMR deployments are expected in the early-to-mid 2030s (NEA, 2025).

Final estimate

For an envelope over the period 2025–2040, we estimate:

$$f_{\text{datacenter}} \approx 20\% \quad (95\%\ \text{CI} : [5\%, 40\%]) \tag{26}$$

Rationale: Since this represents an envelope (upper bound on available power) over the full period, the estimate is weighted toward the later years when datacenter capacity will be highest. The lower bound (5%) reflects near-term constraints (2025–2027). The central estimate (20%) represents mid-to-late period capacity (∼2032–2035). The upper bound (40%) captures the optimistic 2040 exponential projection (37%), with some allowance for faster-than-expected infrastructure buildout.

Estimating f_{AI}

A recent analysis (Vries-Gao, 2025) estimates that "AI systems could be responsible for almost half of data center power demand by the end of 2025."

Final estimate

$$f_{\text{AI}} \approx 70\% \quad (95\%\ \text{CI} : [50\%, 90\%]) \tag{27}$$

Rationale: Assuming continued growth in AI's share of datacenter power (was ∼20% in early 2025, could reach ∼50% by end of 2025), whilst accounting for a practical upper limit due to continued presence of non-AI workloads (traditional cloud, storage, etc.). The lower bound (50%) reflects the 2025 baseline. The central estimate (70%) represents continued growth with some saturation. The upper bound (90%) captures an aggressive scenario where AI dominates datacenter usage.

Estimating f_{training}

According to Moss (2024), OpenAI's training and inference costs could reach $7 billion for 2024, with $4 billion on inference and $3 billion on training. This suggests a current split of

roughly 43% training / 57% inference, likely to skew more toward inference in the future.

However, Epoch AI's analysis of the training-inference compute tradeoff (Erdil, 2024) suggests that comparable investments in training and inference are cost-optimal, so training fractions are unlikely to drop dramatically below $\sim$43%.

Final estimate

$$f_{\text{training}} \approx 37.5\% \quad (95\% \text{ CI} : [20\%, 55\%]) \tag{28}$$

Rationale: The central estimate (37.5%) is slightly below the current observed 43% to account for expected growth in inference demand, while acknowledging that optimal compute allocation theory suggests training fractions should remain substantial. The lower bound (20%) reflects a scenario where inference dominates but training remains economically significant. The upper bound (55%) allows for cases where large training runs consume the majority of AI compute.

Estimating $n_{\text{hyperscalers}}$

There are currently five established frontier model developers actively training large-scale models: xAI, Meta, OpenAI, Google, and Anthropic. Additionally, there are a couple of uncertain newcomers with unclear scaling requirements, such as Thinking Machines and Safe Superintelligence (SSI).

Given the substantial financial opportunity in AI, there is a non-insignificant chance of new entrants over the next 10 years. However, the capital requirements for continued scaling are immense, and funding may become increasingly concentrated among incumbents who have established infrastructure, talent pipelines, and revenue streams. This creates a bottleneck that limits the realistic number of players who can compete at the frontier.

Final estimate

$$n_{\text{hyperscalers}} \sim \text{Uniform}(\{6, 7, 8\}) \tag{29}$$

Rationale: The estimate assumes 5 established developers plus 1–3 additional entrants (either current uncertain players scaling up or new entrants). The discrete uniform distribution over $\{6, 7, 8\}$ reflects uncertainty about whether newcomers will successfully scale to frontier-level training while acknowledging that the field is unlikely to fragment significantly given capital constraints.

Combined Monte Carlo Simulation

Combining the above estimates via Monte Carlo simulation (100,000 samples) yields the following distribution for the overall share:

$$\text{share} = \frac{f_{\text{datacenter}} \cdot f_{\text{AI}} \cdot f_{\text{training}}}{n_{\text{hyperscalers}}} \tag{30}$$

The fraction parameters ($f_{\text{datacenter}}$, f_{AI}, f_{training}) are modelled using truncated normal distributions bounded between 0 and 1 to ensure valid probability values. The number of hyperscalers is sampled from a discrete uniform distribution over $\{6, 7, 8\}$.

The simulation yields a median of 0.71% with a 50% CI of [0.47%, 1.00%] and a 95% CI of [0.13%, 1.74%].

Prior Specification

These fraction parameters encode real-world constraints on power allocation derived from external domain knowledge about the industry. In the full Bayesian model (see main Methods section), these priors remain fixed during hyperparameter optimization, as they represent external constraints rather than relationships to be learned from observational data.

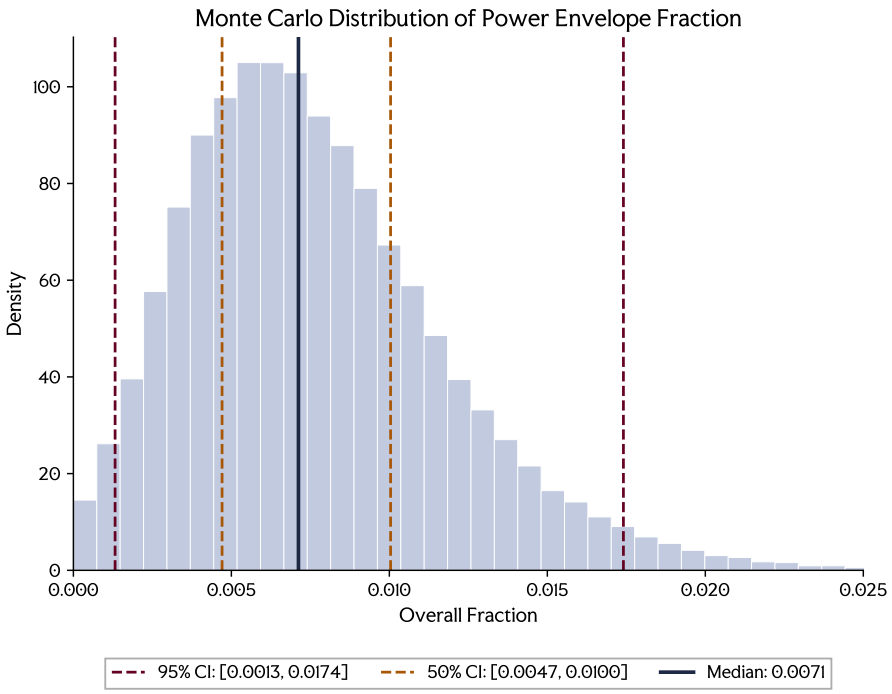

Figure S.1: Monte Carlo distribution of the power share with 50% and 95% confidence intervals marked.

Task Automation Model Fit

Having fitted the task automation model (Section 2.1.3; Equation 1), we evaluate the quality of the fit. The reliability of the task automation model was evaluated by comparing its internal probability estimates against the actual success rates observed in the benchmark data. To do so, we evaluated the alignment between predicted probabilities and empirical outcomes using a binning-based calibration analysis.

For each observation, we calculated the mean predicted probability, $\bar{p}$, across the posterior distribution samples. These probabilities were partitioned into ($n = 10$) equally spaced bins within the $[0, 1]$ interval. Within each bin j, we computed the average predicted probability vs. the observed frequency of outcomes:

$$\text{Observed Frequency}_j = \frac{1}{|B_j|} \sum_{i \in B_j} y_i \tag{31}$$

where B_j is the set of indices of observations falling into bin j, and $y_i \in \{0, 1\}$ is the observed outcome. This relationship was visualized via a reliability diagram (Figure S.2; proximity to the identity line indicates a well-calibrated model). To account for varying data density, marker sizes in the calibration plot were scaled proportional to the number of observations $|B_j|$ in each bin.

The analysis shows that the model's predictions align with the empirical outcomes across the probability spectrum. With an overall accuracy of 0.877 (F1: 0.884, Precision: 0.896, Recall: 0.871), the model maintains a consistent relationship between its confidence levels.

Complementing this, we performed a Posterior Predictive Check (PPC) to determine if the model could statistically reproduce the observed data statistics. We generated S synthetic datasets by drawing binary samples from the posterior predictive distribution:

$$y_{rep,s} \sim \text{Bernoulli}(p_s) \tag{32}$$

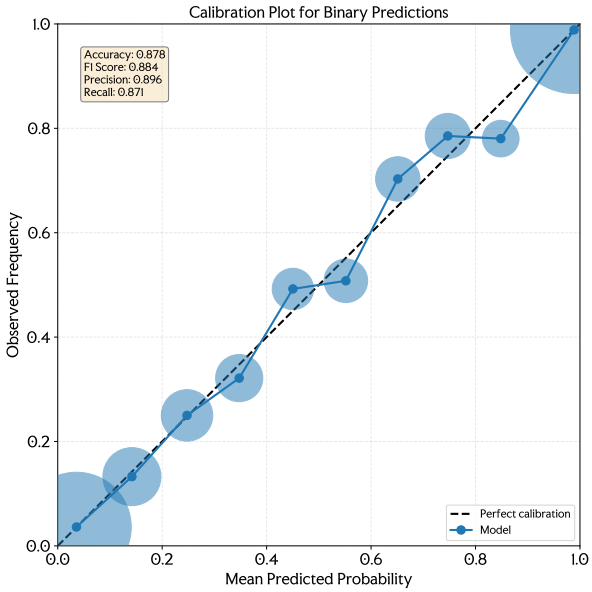

Figure S.2: Reliability of Task Automation Predictions. Observed success frequencies align with the model's predicted probabilities across all probability ranges. Marker sizes are proportional to the number of tasks in each bin; points falling on the diagonal dashed line represent perfect calibration. Summary metrics (inset) confirm the model's overall predictive accuracy and consistency with the benchmark data.

where p_s represents the s-th sample from the posterior distribution of probabilities. We compared the distribution of success rates from these synthetic datasets to the empirical mean success rate. A two-tailed posterior predictive p-value was calculated to assess the consistency between the model and the data. We found strong alignment between the simulated and actual success rates indicates that the model adequately captures the underlying data-generating process (PPC p-value: 0.964; two-sided).

ECI Capability Fits across Algorithmic Improvement Scenarios

Table S.1: ECI capability model fit comparison across algorithmic efficiency functional forms. Coverage = % of observations within 95% credible interval.

	Sigmoid	**Exponential**	**Sum-Sigmoid**
Capability Score ($n = 13$)			
RMSE	5.40	5.32	5.42
MAE	4.35	4.25	4.36
Coverage 95%	92.3%	92.3%	92.3%

Differences to METR Framework Fit

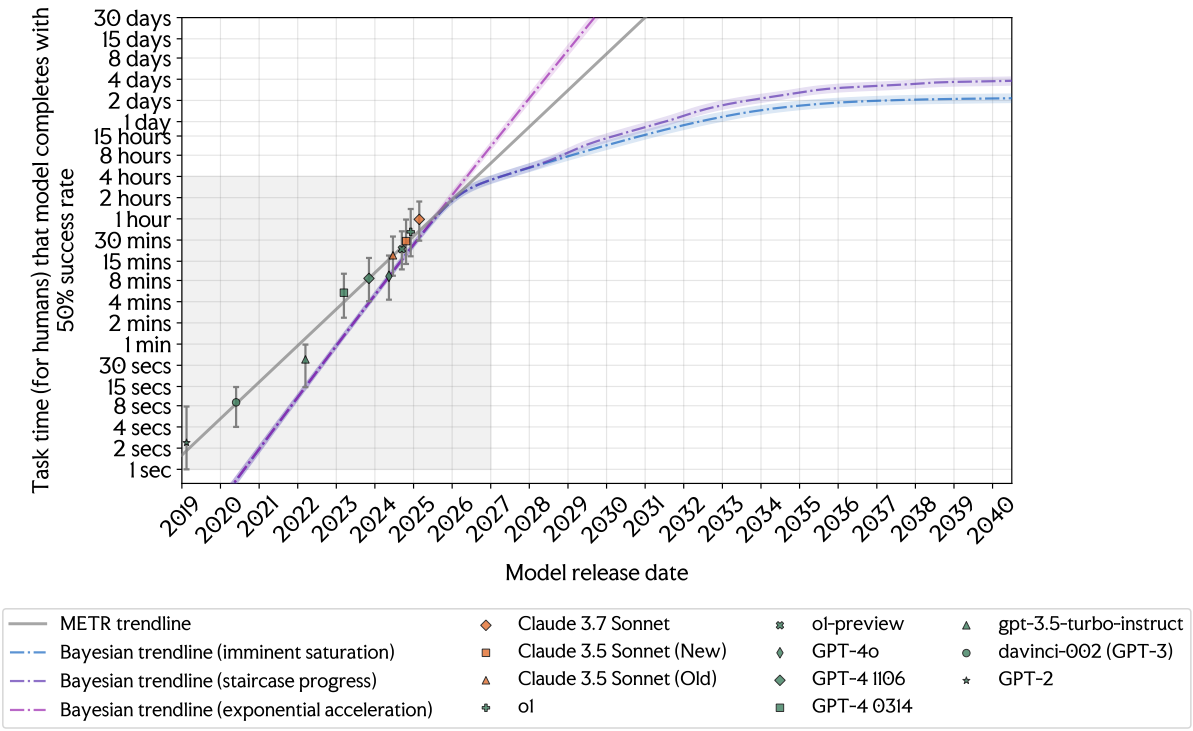

Figure S.3: Forecasted AI task automation capabilities: Bayesian framework vs. METR extrapolation. The shaded region replicates METR's methodology: for each model, a logistic regression predicts task success rates from task length (human completion time), the 50% time horizon is extracted (task length at 50% success, with error bars showing 2.5–97.5% confidence intervals from bootstrapping), and a trendline (gray) is fitted across models to project future capabilities. In contrast, our Bayesian framework jointly models task outcomes as functions of task complexity, task length, and training compute—themselves constrained by power capacity limits and algorithmic efficiency. We show three algorithmic-efficiency scenarios for tasks at average task complexity (messiness = 3.2) from the METR benchmark: imminent saturation (sigmoid; blue dash-dot) yields substantial slowdown and an early plateau; staircase progress (sum of shifted sigmoids; purple dash-dot) allows staircase-shaped growth until progress eventually bottlenecked by power limits; exponential acceleration (violet dash-dot) projects unbounded efficiency gains so that power does not become a bottleneck even by 2040 and beyond. See Results for more details.