
Red Lines Forecasting: When Will Frontier AI Cross Compute Thresholds?

Kwesi Amanfu
Independent

With
Apart Research

Abstract

We forecast when frontier AI training runs will cross compute thresholds relevant to AI governance. Frontier models are those trained with the largest computational budgets each year; we measure training compute in FLOPs (floating-point operations, the total computational work used to train a model). Using Epoch AI data, we find that both policy-defined thresholds (EU AI Act 10^{25} FLOP, US EO 10^{26} FLOP) have already been crossed. Compute milestones at $10\times$, $100\times$, and approximately $3000\times$ the US EO threshold, which are expected within one to three years. We validate forecasts with rolling origin cross-validation and calibration correction. We distinguish policy thresholds from compute milestones. The latter are training budget forecasts, not capability claims.

NOTE: This work was completed over 36 hours as part of a hackathon and the claims require additional validation.

Contents

1	Introduction	3
2	Related Work	3
3	Methods	3
3.1	Data Sources	3
3.2	Threshold Taxonomy	3
3.3	Forecasting Approach	4
3.4	Validation Framework	4
4	Results	4
4.1	Timeline of Threshold Crossings	5
4.2	Capability Forecast Validation	6
4.3	Model Comparison	7
4.4	Calibration Analysis	7
4.5	Backtest Diagnostics	8
4.6	Robustness Checks	8
5	Discussion and Limitations	8
5.1	Implications	8
5.2	Why We Avoid the Capability to Compute Mappings	8
5.3	Limitations	9
5.4	Future Work	9
6	Conclusion	9
7	Code and Data	9
8	References	9
9	Appendix	10
10	LLM Usage Statement	10

1 Introduction

AI governance frameworks increasingly rely on compute thresholds to trigger regulatory requirements. The EU AI Act flags models with more than 10^{25} FLOPs, triggering systemic risk assessment requirements. US Executive Order 14110 (now rescinded) previously required reporting above 10^{26} FLOP. We retain this as a reference threshold given its influence on AI governance discourse.

Policymakers set thresholds but lack tools to forecast when they will be crossed. This creates a gap between regulatory intent and practical implementation. Our work addresses this gap by providing empirically grounded forecasts of when frontier training runs will cross these thresholds and milestones.

The Epoch Capabilities Index (ECI) aggregates performance across major AI benchmarks into a single score. Epoch defines 165 ECI as roughly PhD-level problem solving, meaning the ability to solve novel research-level problems. We use ECI to cross-validate our compute forecasts with capability trends.

Main Contributions

Our main contributions are:

1. Empirical forecasts showing both policy-defined thresholds (EU AI Act, US EO) have already been crossed, with future milestones arriving within one to three years
2. A rigorous validation framework using rolling origin cross-validation and calibration correction
3. Conceptual distinction between policy thresholds (which trigger regulations) and compute milestones (which are forecasting markers without regulatory weight)

2 Related Work

Sevilla et al. (2022) established the foundational analysis of compute trends across three eras of machine learning, demonstrating exponential growth in training compute over time. Our work extends this by focusing specifically on governance-relevant thresholds and providing uncertainty-quantified forecasts.

Erben et al. (2025) analyzed training compute thresholds for the EU AI Act and found that compute thresholds only weakly correlate with aggregated task performance. This finding motivates our careful distinction between compute milestones and capability claims.

Anthropic’s Responsible Scaling Policy (2023) defines AI Safety Levels based on capability evaluations, not on compute thresholds. This approach recognizes the limitations of FLOP as a proxy for capability and informs our methodological choices.

Our work differs from prior analyses by providing explicit timeline forecasts with calibrated uncertainty intervals, validated through backtesting rather than solely retrospective trend analysis.

3 Methods

3.1 Data Sources

3.2 Threshold Taxonomy

We distinguish two categories of thresholds:

Policy Defined Thresholds (solid lines in figures): These appear explicitly in regulatory text and trigger actual governance mechanisms.

Source	Coverage	Usage
Epoch AI Notable Models	2012–2025	Annual frontier compute
Epoch Capabilities Index	2023–2025	Monthly frontier capability

Table 1: Data sources used in this study

- EU AI Act: 10^{25} FLOP
- US EO 14110: 10^{26} FLOP (now rescinded but retained as reference)

Compute Milestones (dotted lines in figures): These are forecasting markers we define for visualization purposes. They carry no regulatory weight and are not capability claims.

- $10\times$ US EO: 10^{27} FLOP
- $100\times$ US EO: 10^{28} FLOP
- Approximately $3000\times$ US EO: $10^{29.5}$ FLOP

3.3 Forecasting Approach

We forecast in $\log_{10}(\text{FLOP})$ space, meaning we model 10^{25} , 10^{26} , 10^{27} rather than the raw numbers which span many orders of magnitude. In this space, a linear trend corresponds to exponential growth in raw FLOP (approximately $4.5\times$ per year).

We compared five models: naive, linear, piecewise, quadratic, and exponential. We selected piecewise for best fit. More complex models do not improve short horizon forecasts ($\Delta\text{BIC} < 2$ versus more complex alternatives).

3.4 Validation Framework

Our validation approach includes:

- Rolling origin cross validation at one, two, and three year horizons
- Calibration correction to achieve approximately 90% prediction interval coverage
- Bootstrap resampling with 2,000 iterations for uncertainty quantification

4 Results

Key Findings

Finding	Timeline	Status
Both policy thresholds crossed	2022–2024	Complete
Next milestone ($10\times$ US EO)	2026	Imminent
Major uncertainty beyond	2027+	95% CI >2 years

Table 2: Summary of main findings

Figure 1 shows our main forecast results. Both policy thresholds have already been crossed. The EU AI Act threshold (approximately 10^{25} FLOP) was crossed around 2022. The US EO threshold (approximately 10^{26} FLOP) was crossed around 2024.

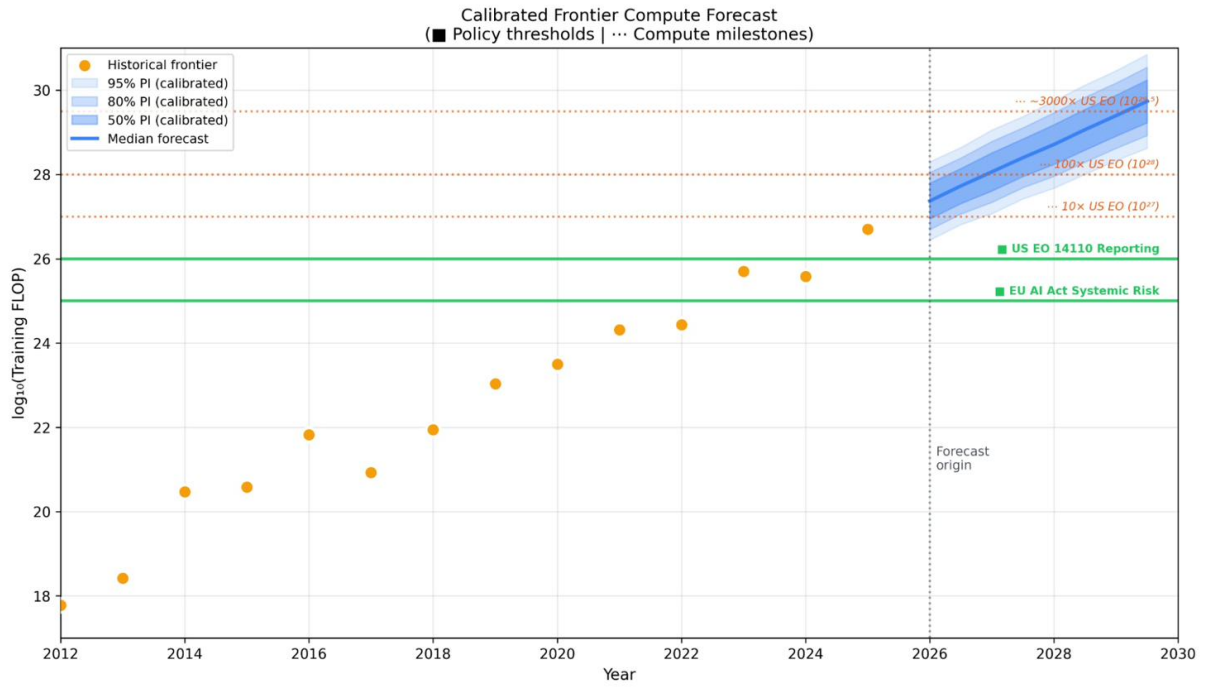


Figure 1: Frontier AI compute forecast (2012–2029). Gold dots represent historical frontier models. Blue shaded regions show 50%, 80%, and 95% prediction intervals. Solid lines indicate policy-defined thresholds (EU AI Act, US EO). Dotted lines indicate compute milestones (10×, 100×, approximately 3000× US EO).

4.1 Timeline of Threshold Crossings

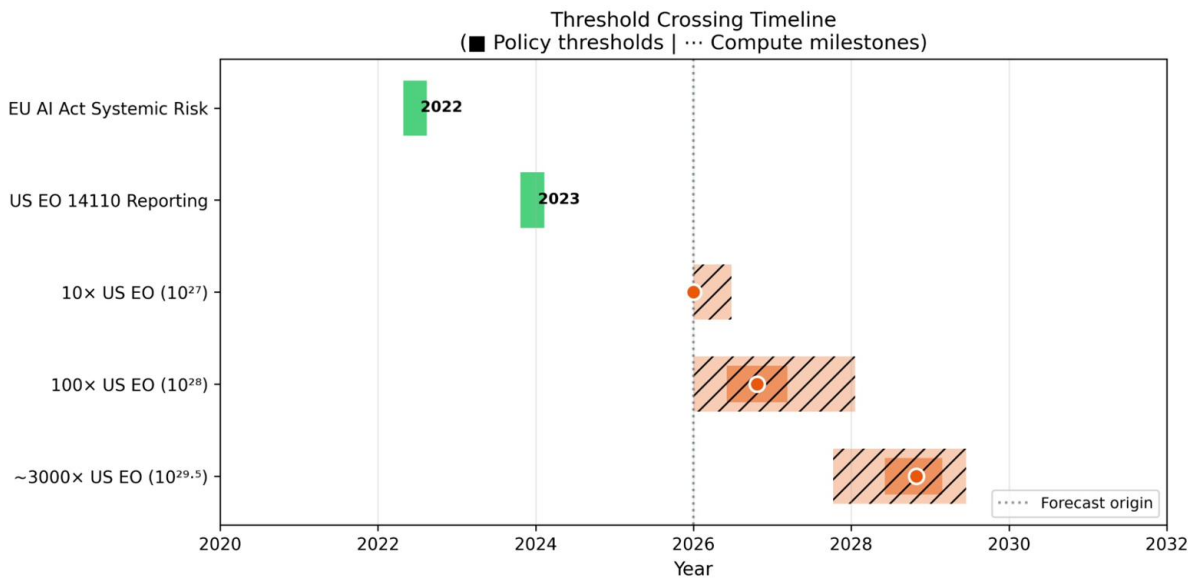


Figure 2: Timeline of threshold crossings. Green indicates already crossed. Orange with hatching indicates future estimates with 50% and 95% confidence intervals.

Threshold	Value	Estimated Crossing	Status
EU AI Act	10^{25} FLOP	Approximately 2022	Crossed
US EO 14110	10^{26} FLOP	Approximately 2024	Crossed
$10\times$ US EO	10^{27} FLOP	2026	Imminent
$100\times$ US EO	10^{28} FLOP	2027	Approximately 1 year
Approximately $3000\times$ US EO	$10^{29.5}$ FLOP	2028–2029	Approximately 2.5 years

Table 3: Summary of threshold crossings

4.2 Capability Forecast Validation

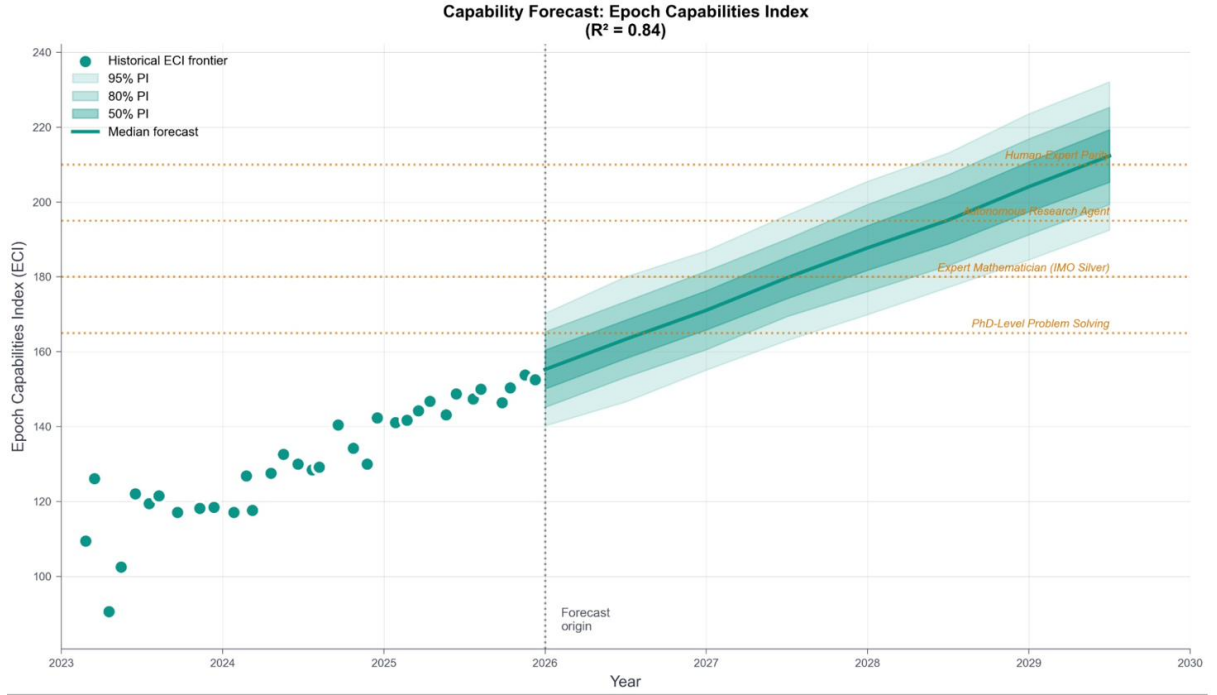


Figure 3: Capability forecast using Epoch Capabilities Index (ECI). The current frontier is approximately 154 ECI. Growth is approximately 16 points per year. The ECI capability forecast predicts PhD-level problem solving (approximately 165 ECI) by late 2026, roughly aligning with the $10\times$ US EO compute milestone.

4.3 Model Comparison

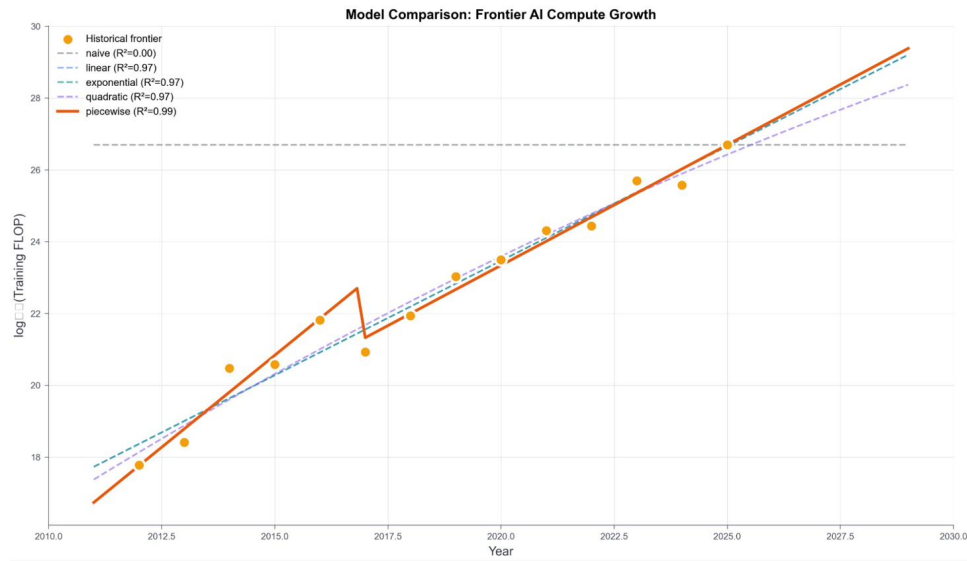


Figure 4: Five models compared. Linear is selected for interpretability, given a ΔBIC of less than 2 relative to more complex models.

4.4 Calibration Analysis

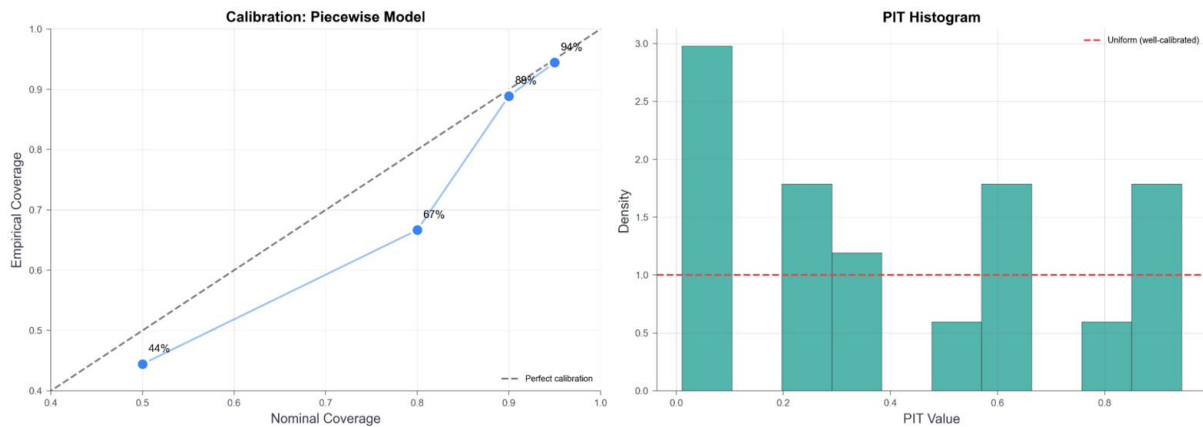


Figure 5: The left panel shows empirical versus nominal coverage. The right panel shows a PIT histogram, where a uniform distribution indicates well-calibrated predictions.

4.5 Backtest Diagnostics

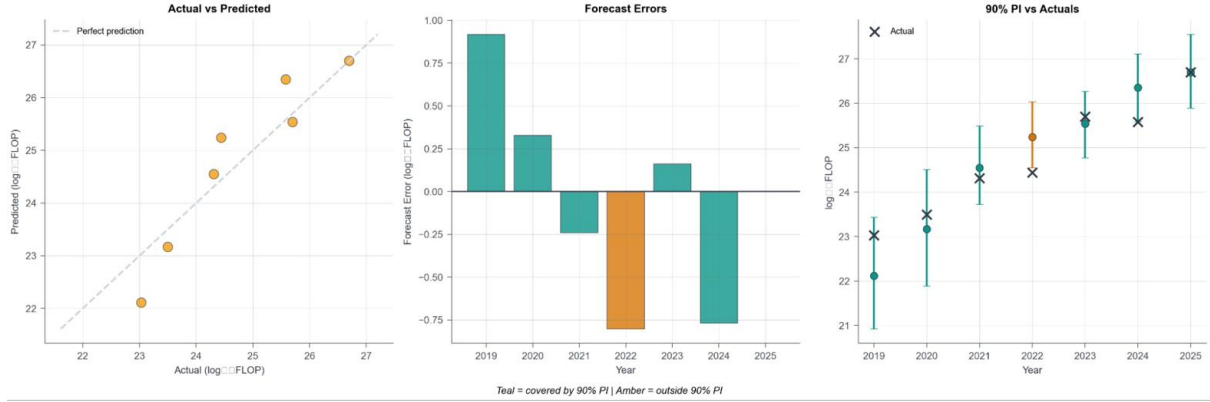


Figure 6: Rolling origin backtest. Teal indicates actual values within 90% prediction interval. Amber indicates values outside the interval.

4.6 Robustness Checks

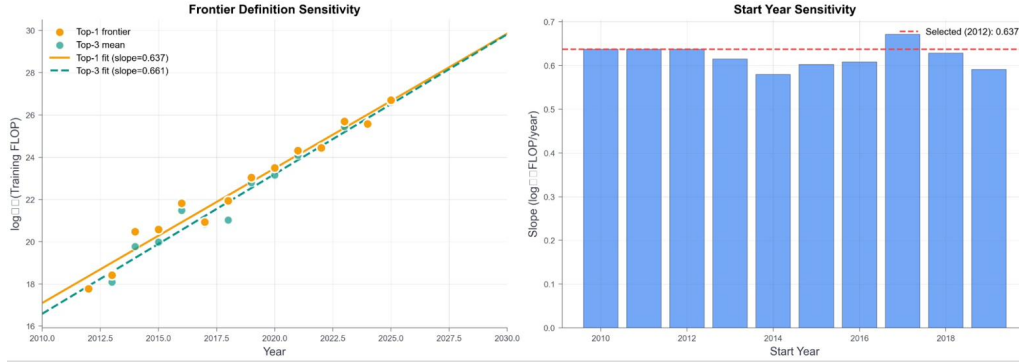


Figure 7: Sensitivity analysis examining frontier definition (top 1 versus top 3) and start year.

5 Discussion and Limitations

5.1 Implications

Our analysis reveals that governance mechanisms should already be in place, as both existing policy thresholds have been exceeded. Future milestones arrive quickly: $10\times$ US EO in 2026, $100\times$ in 2027. Uncertainty grows rapidly, with the 95% confidence interval spanning more than two years for milestones beyond 2027.

These findings suggest that policymakers should treat current thresholds as already binding while preparing governance frameworks for the next generation of frontier models.

5.2 Why We Avoid the Capability to Compute Mappings

ASL 4 (AI Safety Level 4) is a capability tier in Anthropic’s Responsible Scaling Policy that triggers safety measures when models can autonomously conduct novel research. Crucially, Anthropic defines ASL levels based on capability evaluations, not FLOPs. There is no “ASL 4 equals X FLOP” formula in the literature. We avoid this common error by using policy-anchored milestones instead.

5.3 Limitations

Compute does not equal capability. We forecast when training budgets will grow, not when capabilities will emerge. As Erben et al. (2025) show, compute only weakly correlates with aggregated task performance. The compute milestones are training budget markers, not capability claims.

Threshold uncertainty. Beyond the EU AI Act and US EO, our milestones are forecasting markers rather than literature-backed claims.

No saturation modeling. There is growing discussion of hitting a wall, meaning frontier labs may face data exhaustion (running out of high-quality training data), diminishing algorithmic returns (each doubling of compute yields smaller capability gains), or hardware limits (chip manufacturing and energy constraints). Our exponential models cannot capture S-curve saturation. Historical data (2012–2025) shows no clear slowdown, but if frontier labs are already seeing diminished returns, our forecasts would be optimistic.

Short-horizon reliability. Uncertainty grows rapidly. The 95% confidence interval extends beyond 2027 by more than 2 years.

5.4 Future Work

Future extensions could incorporate trends in algorithmic efficiency to model effective compute rather than raw FLOPs. Saturation modeling using S-curve or logistic growth models could address potential slowdowns. Integration with capability evaluations could provide more direct governance relevance.

6 Conclusion

Both policy-defined thresholds (EU AI Act 10²⁵, US EO 10²⁶) have been crossed. Compute milestones at 10×, 100×, and approximately 3000× the US EO threshold are expected within one to three years under current trends.

These are training budget forecasts, not capability predictions. Governance mechanisms triggered at these compute levels should incorporate capability evaluations rather than rely solely on FLOP.

7 Code and Data

- **Code repository:** https://github.com/retroam/apart/tree/main/technical_ai_governance
- **Data:** Epoch AI Notable Models, Epoch AI Benchmarking Hub (both CC BY)
- **Notebook:** `redline_forecast_submission.ipynb`
- **Outputs:** `output/` directory

8 References

1. Epoch AI. (2025). Notable AI Models Database. <https://epoch.ai/data/notable-ai-models>
2. Epoch AI. (2025). AI Benchmarking Hub and Epoch Capabilities Index. <https://epoch.ai/benchmarks>

3. Erben, A., Negele, M., Heim, L., Sevilla, J., Fernandez Llorca, D., and Gomez, E. (2025). Training Compute Thresholds: Key Considerations for the EU AI Act. JRC Publications Repository. <https://publications.jrc.ec.europa.eu/repository/handle/JRC143255>
4. European Union. (2024). AI Act, Article 51. Official Journal of the European Union.
5. Executive Office of the President. (2023). Executive Order 14110. Federal Register. (Rescinded January 20, 2025, by EO 14148.)
6. Sevilla, J., et al. (2022). Compute Trends Across Three Eras of Machine Learning. arXiv:2202.05924.
7. Anthropic. (2023). Anthropic’s Responsible Scaling Policy. (Defines ASL via capability evaluations, not computation.)

9 Appendix

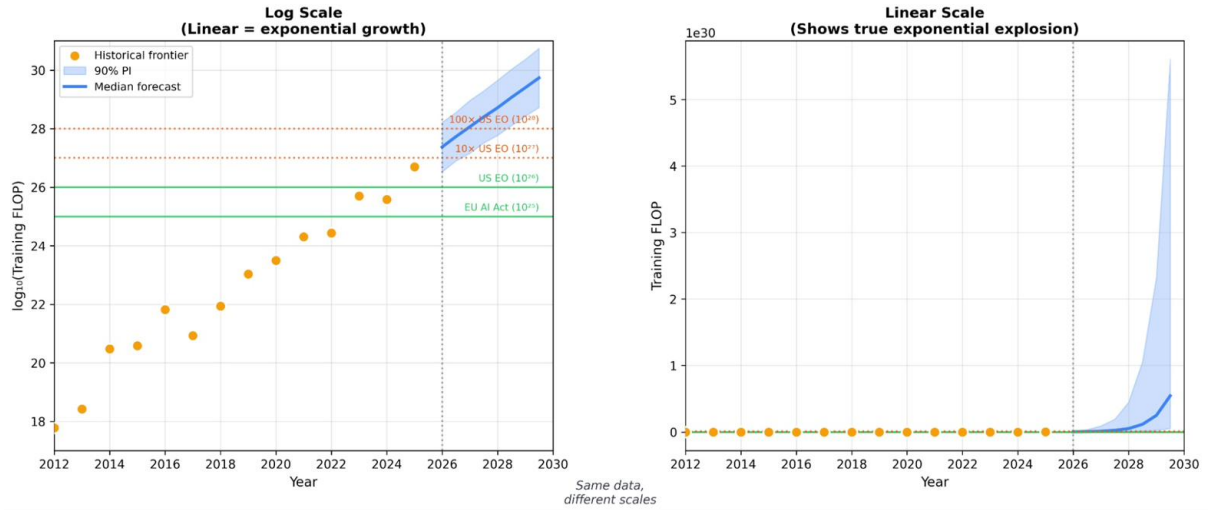


Figure 8: Log scale versus linear scale comparison. The left panel shows the $\log_{10}$ view we use for modeling, where a linear trend corresponds to exponential growth. The right panel shows the raw FLOP, demonstrating the true exponential explosion. All historical data and policy thresholds appear near zero relative to 2030 forecasts.

10 LLM Usage Statement

Claude assisted with code development, and OpenAI Prism was used in drafting. All results were verified manually.