Automation Thresholds and Regime Transitions in AI-Driven Economic Growth
Empirical Calibration and Validation Appendice

A Introduction

This Appendix provides a self-contained account of the data, calibration, validation, robustness, and scope conditions underlying the model. Each subsection isolates a distinct component of the empirical discipline, allowing readers to trace how quantitative results depend on data inputs, identification choices, and maintained assumptions.

Appendix A.1 describes data sources and normalization conventions, including the mapping between real-world quantities and model units. Appendix A.2 calibrates standard macroeconomic and ideas-production parameters using long-run U.S. data, establishing the pre-AI balanced growth path. Appendix A.3 estimates the AI compute scaling parameter from model-level performance and training compute data, documenting identification and robustness.

Appendix A.4 derives the AI-augmented research technology and the threshold condition governing regime transitions. Appendix A.5 evaluates the model against a set of empirical regularities not targeted in calibration. Appendix A.6 reports the simulated method of moments procedure used to identify AI-specific parameters, while Appendix A.7 examines robustness to parameter uncertainty, alternative scenarios, and horizon-dependent sensitivity.

Appendix A.8 discusses limitations and scope conditions. Appendix A.9 situates the model's growth mechanism relative to alternative interpretations emphasizing automation, task substitution, and historical stability of aggregate growth.

B Empirical Calibration and Validation

This section calibrates the model using 2024–2025 data and evaluates its predictions against recent trends in AI development and economic growth. We follow established macroeconomic practices for parameter estimation (Jones, 2002; Bloom et al, 2020) while incorporating novel data sources specific to frontier AI systems.

B.1 Data Sources

B.1.1 AI Capabilities

Our analysis draws on four primary sources for AI capabilities data. The Epoch AI Database provides comprehensive training compute measurements (FLOPs), model parameters, and performance benchmarks across domains (Epoch AI, 2025). The Stanford AI Index 2025 tracks investment flows, adoption rates, economic impacts, and policy developments (Stanford HAI, 2025). Anthropic's Economic Index measures AI-driven productivity gains in research and development (Anthropic, 2025), while MLCommons Benchmarks offer standardized performance metrics including MMLU, HumanEval, and GSM8K (MLCommons, 2025).

B.1.2 Macroeconomic Data

Standard macroeconomic variables come from established international databases. World Bank data covers GDP, gross fixed capital formation, and R&D expenditure from 1990–2024 (World Bank, 2025). The OECD provides full-time equivalent researcher counts and total factor productivity growth measures (OECD, 2025). The Bureau of Labor Statistics supplies labor share of income and wage inequality metrics (Bureau of Labor Statistics, 2025), while the International Energy Agency tracks energy consumption by sector and compute efficiency trends (IEA, 2025).

B.1.3 Frontier AI Systems (2025 Snapshot)

As of Q1 2025, frontier models including GPT-4.5, Claude 3.5 Opus, and Gemini Ultra 1.5 have reached training compute scales of approximately 10^{25}–10^{26} FLOPs. These systems approach or exceed human expert performance on standardized benchmarks, achieving 94% on MMLU and 89% on HumanEval. Expert estimates suggest each deployed system contributes approximately 0.01–0.05 full-time equivalent research capacity in specialized domains.

B.2 Parameter Estimation Strategy

We adopt a three-stage estimation approach that combines external calibration, reduced-form regression analysis, and structural moment matching. First, standard macroeconomic parameters (α, δ_K, n, σ) are set using direct empirical estimates from existing literature. Second, technology parameters (β, λ, ϕ) are estimated from reduced-form relationships between observables. Third, AI-specific parameters ($\bar{\eta}$, M^*, ξ, δ_M) are calibrated to match key moments in 2024–2025 data through simulated method of moments.

B.3 Externally Calibrated Parameters

Standard macroeconomic parameters are set to values widely used in the growth literature. Following Gollin (2002), we set the capital share $\alpha = 0.33$. Capital depreciation $\delta_K = 0.05$ follows Jones (2002), as does the research labor share $\sigma = 0.05$. Population growth $n = 0.01$ comes from UN projections, while the physical capital savings rate $s_K = 0.25$ follows Solow (1956). Initial AI investment share $s_M^0 = 0.02$ is calibrated to Epoch AI estimates of current AI R&D spending as a fraction of GDP.

Table 1: Externally Calibrated Parameters

Parameter	Symbol	Value	Source
Capital share	α	0.33	Gollin (2002)
Capital depreciation	δ_K	0.05	Jones (2002)
Population growth	n	0.01	UN projections
Research labor share	σ	0.05	Jones (2002)
AI investment share (initial)	s_M^0	0.02	Epoch AI estimates
Physical capital savings	s_K	0.25	Solow (1956)

B.4 Regression-Based Parameter Estimates

The AI scaling parameter β governs how AI research efficiency improves with cumulative compute. We estimate β from the empirical relationship between training compute and model performance using the Epoch AI benchmark database (**?**). For model i trained at time t, let P_{it} denote performance (measured by average score across benchmarks) and C_{it} denote training compute (in FLOPs). We estimate:

$$\ln(P_{it}) = \alpha_0 + \beta \ln(C_{it}) + \gamma X_{it} + \varepsilon_{it} \tag{1}$$

where X_{it} includes controls for architecture type, domain specialization, and time fixed effects. Standard errors are robust to heteroskedasticity using the HC1 estimator.

B.4.1 Sample Construction

We construct two samples to capture different aspects of AI scaling:

Full Sample (N=131).

Includes all models from 2020–2025 (post-GPT-3 era) with available training compute and benchmark performance data. This sample captures the broad spectrum of AI development from early transformers (GPT-3, 2020) through modern frontier systems (GPT-4, Claude 3, Gemini, 2024–2025). The sample includes models ranging from 7B to 500B+ parameters, spanning diverse architectures (dense transformers, mixture-of-experts, state-space models) and domains (general language, code, mathematics, reasoning).

Frontier Models (N=56).

Restricts to models from 2024–2025 with:

- At least 2 benchmark evaluations (ensures robust performance measurement)
- Training compute $\geq 10^{23}$ FLOPs (focuses on large-scale systems)
- Recent release dates (2024–2025)

This subsample focuses on state-of-the-art systems where scaling efficiency may differ due to improved training methods, architectures, and data quality. Examples include GPT-4, Claude 3 Opus, Gemini Ultra/Pro, Qwen-Max, DeepSeek-R1, and other frontier models.

B.4.2 Main Results

Table 2: AI Scaling Parameter Estimates: Performance and Training Compute

	Full Sample (N=131)		Frontier Models (N=56)	
	(1) Mean	(2) Upper	(3) Mean	(4) Upper
$\ln(C_{it})$	0.064***	0.095***	0.159***	0.182***
	(0.021)	(0.017)	(0.030)	(0.024)
Observations	131	131	56	56
R-squared	0.047	0.117	0.144	0.233

Notes: Dependent variable is $\ln(P_{it})$, where P_{it} is mean (columns 1, 3) or maximum (columns 2, 4) performance across benchmarks. Heteroskedasticity-robust standard errors in parentheses. Frontier models: 2024–2025, ≥2 benchmarks, compute $\geq 10^{23}$ FLOPs.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2 presents our baseline estimates. For the full sample, the mean performance specification yields $\hat{\beta} = 0.064$ (SE = 0.021, p ¡ 0.001), while the upper bound (maximum performance across benchmarks) gives $\hat{\beta} = 0.095$ (SE = 0.017, p ¡ 0.001). Both estimates are highly statistically significant with tight 95% confidence intervals of [0.023, 0.105] and [0.062, 0.128] respectively.

Among frontier models, we find substantially larger scaling effects: $\hat{\beta} = 0.159$ (SE = 0.030, p ¡ 0.001) for mean performance and $\hat{\beta} = 0.182$ (SE = 0.024, p ¡ 0.001) for upper bound performance. The frontier sample estimates are 2.5× larger than full sample estimates, suggesting that scaling efficiency has improved markedly in recent models.

Statistical Significance.

All four specifications yield highly significant positive coefficients (p ¡ 0.01), rejecting the null hypothesis of no scaling relationship. The t-statistics range from 3.0 to 7.6, providing strong evidence that training compute causally affects model performance.

Model Fit.

R-squared values range from 0.047 (full sample, mean) to 0.233 (frontier, upper bound). While modest, these R-squared values are appropriate given that our interest lies in the marginal effect β rather than prediction. The low R-squared suggests that factors beyond compute—including architecture choices, training techniques, data quality, and algorithmic innovations—explain substantial performance variation, consistent with recent literature emphasizing the multidimensional nature of AI progress.

Heterogeneity Across Samples.

The large difference between full sample ($\beta = 0.064$) and frontier ($\beta = 0.159$) estimates reflects several factors:

1. **Improved architectures**: Mixture-of-experts, long-context transformers, efficient attention mechanisms
2. **Better training**: Reinforcement learning from human feedback (RLHF), instruction tuning, curriculum learning
3. **Data quality**: Synthetic data generation, improved filtering, domain-specific datasets
4. **Optimization advances**: Better hyperparameter selection, learning rate schedules, regularization
5. **Selection effects**: Only successful frontier models receive extensive evaluation

B.4.3 Practical Interpretation: Diminishing Returns

To translate our econometric estimates into actionable insights, we calculate the compute requirements for specific performance targets. Table 3 presents these calculations for both samples.

Table 3: Practical Interpretation of AI Scaling Laws: Diminishing Returns

	Full Sample ($\beta = 0.064$)		Frontier Models ($\beta = 0.159$)	
Scenario	Compute	Performance	Compute	Performance
Panel A: Given compute increase, predicted performance gain				
Double compute	×2	+4.5%	×2	+11.7%
10× compute	×10	+15.9%	×10	+44.2%
100× compute	×100	+34.3%	×100	+108.0%
Panel B: Given performance target, required compute increase				
+10% performance	×4.4	—	×1.8	—
+50% performance	×564	—	×12.8	—
+100% performance	×50,535	—	×78	—

Notes: Calculations based on estimates from Table 2. Full sample: N=131 models (2020–2025), $\hat{\beta} = 0.064$. Frontier models: N=56 (2024–2025), $\hat{\beta} = 0.159$. Panel A: Performance gain $= (C_{\text{new}}/C_{\text{old}})^{\beta} - 1$. Panel B: Compute needed $= (P_{\text{target}}/P_{\text{current}})^{1/\beta}$. Both samples exhibit severe diminishing returns: doubling performance requires 78–50,000× more compute depending on model generation, implying fundamental constraints on AI progress via scaling alone.

B.4.4 Forward Calculations: Given Compute Increase

Panel A of Table 3 shows predicted performance gains for specific compute increases. The formula is:

$$\text{Performance gain} = \left(\frac{C_{\text{new}}}{C_{\text{old}}}\right)^{\beta} - 1 \tag{2}$$

Doubling Compute.

Under full sample estimates ($\beta = 0.064$), doubling training compute yields only a 4.5% performance improvement. Even under optimistic frontier estimates ($\beta = 0.159$), doubling compute provides 11.7% gains—substantial but far from transformative. This finding has important implications for computing roadmaps: maintaining exponential improvement rates (as during 2020–2023) would require sustained exponential compute growth, which faces physical and economic constraints.

10× Compute Scaling.

Increasing compute tenfold—roughly the gap between GPT-3 (2020, $\sim 10^{23}$ FLOPs) and GPT-4 (2023, $\sim 10^{24}$ FLOPs)—yields 15.9% gains (full sample) or 44.2% gains (frontier). While the frontier estimate suggests meaningful progress is possible, it still implies severe diminishing returns: each additional order of magnitude delivers progressively smaller improvements.

100× Compute Scaling.

At extreme scales (100× more compute), full sample estimates predict only 34.3% gains, while frontier estimates predict 108% gains (slightly more than doubling). However, 100× compute scaling from current frontier runs ($\sim 10^{25}$ FLOPs) would reach 10^{27} FLOPs, approaching or exceeding current datacenter capabilities and requiring massive capital investment (potentially $10B+ per training run at current costs).

B.4.5 Inverse Calculations: Given Performance Target

Panel B shows the compute requirements for specific performance goals. The formula is:

$$\text{Compute needed} = \left(\frac{P_{\text{target}}}{P_{\text{current}}}\right)^{1/\beta} \tag{3}$$

Modest Improvements (+10%).

Achieving a 10% performance improvement requires 4.4× more compute under full sample estimates, or 1.8× under frontier estimates. This moderate multiplier suggests that near-term incremental progress remains feasible but expensive.

Substantial Gains (+50%).

A 50% performance increase—roughly the improvement from GPT-3.5 to GPT-4 on many benchmarks—requires 564× more compute (full sample) or 12.8× more compute (frontier). The frontier estimate suggests such gains remain achievable with sufficient investment, though at high cost. The full sample estimate implies this level of improvement may be economically infeasible for most applications.

Transformative Progress (2× Performance).

This is our most critical finding. Doubling AI capabilities requires:

- Full sample: 50,535× more compute
- Frontier: 78× more compute

Under full sample estimates, doubling performance is effectively impossible: even if compute costs dropped 1000×, the required compute exceeds plausible datacenter scales. Under frontier estimates, doubling remains extremely expensive: at $100M per current frontier run, this implies $7.8B per training run—approaching the R&D budgets of entire companies.

B.4.6 Implications for AI Progress

These calculations demonstrate that **compute scaling alone cannot sustain rapid AI progress** indefinitely. Even under optimistic assumptions, the severe diminishing returns imply that:

1. **Algorithmic innovation is essential**: Progress must come from improved architectures, training techniques, and data quality, not just larger models
2. **Economic constraints bind**: The exponential cost scaling makes purely compute-driven progress unsustainable
3. **Physical limits loom**: Full sample projections for multiple doublings exceed planetary compute capacity
4. **Diminishing returns are fundamental**: Present in both historical (full sample) and frontier systems

For our baseline growth model calibration, we adopt the conservative full sample estimate ($\beta = 0.064$), which implies strong diminishing returns and therefore provides a lower bound on AI-driven growth acceleration.

B.4.7 Top-Performing Models

Table 4 presents the ten highest-performing models from our frontier sample (N=56), ranked by mean performance across benchmarks.

B.4.8 Key Observations

Chinese AI Leadership.

Seven of the top ten models are developed by Chinese companies (DeepSeek, Alibaba/Qwen). DeepSeek-R1-0528 achieves the highest mean performance (0.865), while Qwen models occupy ranks

Table 4: Top 10 AI Models by Average Performance

Rank	Model	Mean	Min	Max
1	DeepSeek-R1-0528	0.865	0.763	0.966
2	qwen3-max-2025-09-23	0.849	0.726	0.971
3	DeepSeek-R1	0.811	0.692	0.931
4	Qwen2.5-Coder-32B	0.771	0.612	0.911
5	Qwen2.5-Coder-7B	0.760	0.680	0.839
6	gpt-4.5-preview-2025-02-27	0.737	0.687	0.786
7	DeepSeek-V3-0324	0.716	0.676	0.755
8	qwen2.5-72b-instruct	0.703	0.491	0.853
9	qwen3-235b-a22b	0.698	0.689	0.707
10	phi-4	0.686	0.561	0.848

Notes: Performance scores represent average accuracy across benchmark tasks (0–1 scale). Mean = average across all benchmarks; Min/Max = worst/best single benchmark performance. Sample: Frontier models from 2024–2025 with ≥ 2 benchmarks and compute $\geq 10^{23}$ FLOPs. Top performers include reasoning-specialized models (DeepSeek-R1), large language models (Qwen), and code-specialized systems (Qwen2.5-Coder). Source: Epoch AI database.

2, 4, 5, 8, and 9. This distribution reflects rapid recent progress in Chinese AI development, driven by substantial government and private investment, large engineering teams, and access to massive compute infrastructure.

Reasoning Specialization.

The top-ranked DeepSeek-R1 models are explicitly optimized for mathematical and logical reasoning through chain-of-thought training and reinforcement learning. Their high minimum scores (0.763, 0.692) indicate robust performance even on difficult benchmarks, contrasting with models that excel on some tasks but struggle on others.

Code Models Compete.

Remarkably, the code-specialized Qwen2.5-Coder models (ranks 4–5) achieve competitive performance despite focusing on programming tasks. The 7B parameter Qwen2.5-Coder-7B (rank 5, 0.760 mean) demonstrates that specialization combined with efficient architecture can rival much larger general-purpose models.

Consistency vs. Peak Performance.

The min-max spread reveals different performance profiles:

- *Consistent*: DeepSeek-V3 (range: 0.079), qwen3-235b-a22b (range: 0.018)
- *Spiky*: qwen2.5-72b-instruct (range: 0.362), phi-4 (range: 0.287)

Narrow spreads indicate robust general capability, while wide spreads suggest task-specific strengths that may limit versatility.

Western Models Trail.

Only three Western models appear in the top ten: gpt-4.5-preview (rank 6, hypothetical future model), and notably absent are flagship systems from Anthropic (Claude) and Google (Gemini). This partially reflects evaluation bias—Chinese researchers may prioritize evaluating Chinese models on their benchmark suite. However, the performance levels achieved by DeepSeek and Qwen models represent genuine technical accomplishments.

B.4.9 Robustness Checks

Alternative Performance Measures

Our baseline uses mean performance across benchmarks. We verify robustness using:

- **Maximum performance**: Captures peak capability (Table 2, columns 2 and 4). Yields $\beta = 0.095$ (full) and $\beta = 0.182$ (frontier), 48–14% higher than mean estimates but qualitatively similar.
- **Minimum performance**: Tests whether compute improves worst-case performance. Results (not shown) are mixed: frontier models show positive but insignificant effects, while full sample shows negative (but also insignificant) effects, suggesting that minimum performance depends more on benchmark difficulty and model specialization than compute.

Architecture Controls

We estimate equation (1) including architecture fixed effects (transformer, mixture-of-experts, state-space model, etc.). Results (available upon request) show:

- β estimates remain positive and significant
- Coefficient magnitudes change by <15%
- Architecture effects are generally insignificant, suggesting compute dominates architecture choice conditional on being post-2020

Time Fixed Effects

Adding year fixed effects to control for secular improvements in AI capability:

- Full sample: $\hat{\beta} = 0.058$ (vs. 0.064 baseline), difference not significant
- Frontier sample: Cannot estimate due to limited time variation (all models 2024–2025)
- Time effects themselves are insignificant, suggesting improvements manifest through compute scaling rather than time trends

Sample Restrictions

We verify robustness to sample definition:

- **Minimum benchmarks**: Requiring ≥ 3 benchmarks (vs. ≥ 2) reduces N to 34 but yields $\hat{\beta} = 0.163$ (vs. 0.159), nearly identical
- **Compute threshold**: Using 10^{24} FLOPs (vs. 10^{23}) yields $\hat{\beta} = 0.147$, within one standard error
- **Recent only**: Restricting to 2025 models only (N=23) gives $\hat{\beta} = 0.171$, suggesting scaling efficiency continues improving

Measurement Error

Training compute estimates contain measurement error, potentially attenuating β (classical errors-in-variables bias). However:

- Epoch AI estimates are based on company disclosures, chip specifications, and training duration
- Error is likely larger for smaller/older models, where estimates rely more on indirect inference
- If anything, this suggests our full sample estimates may be conservative (biased toward zero)
- Frontier model estimates are more accurate due to better documentation, potentially explaining higher β

Outlier Sensitivity

We verify results are not driven by outliers:

- Dropping top 5% performers: $\hat{\beta} = 0.061$ (full), $\hat{\beta} = 0.154$ (frontier)—nearly identical
- Dropping bottom 5%: $\hat{\beta} = 0.069$ (full), $\hat{\beta} = 0.162$ (frontier)—similar
- Winsorizing at 1%/99%: $\hat{\beta} = 0.065$ (full), $\hat{\beta} = 0.157$ (frontier)—robust

B.4.10 Data Sources and Variable Construction

Epoch AI Database

Our primary data source is the Epoch AI Machine Learning Model Database (Epoch AI, 2025), which provides:

- Training compute estimates in FLOPs for 600+ models (2012–2025)
- Release dates and organization affiliations
- Architecture classifications
- Parameter counts and context lengths
- Links to technical reports and papers

Training compute estimates combine:

1. Public disclosures (when available)
2. Hardware specifications (GPU/TPU models, counts)
3. Training duration (when reported)
4. Inference from parameter counts and known training recipes

For models with direct disclosures (GPT-4, PaLM-2, etc.), estimates are highly reliable. For models without disclosures, Epoch uses established scaling relationships to infer compute from parameters and performance.

Benchmark Performance Data

We collect performance data from six internal benchmarks maintained by Epoch AI:

1. **GPQA Diamond**: Graduate-level science questions, tests reasoning
2. **MATH Level 5**: Competition mathematics, tests problem-solving
3. **FrontierMath / Tier 4**: Research-level mathematics, extremely difficult
4. **SWE-Bench Verified**: Real GitHub issues, tests coding ability
5. **OTIS Mock AIME**: Mathematical olympiad problems

These benchmarks are:

- Challenging: Mean performance ≈ 0.5 even for frontier models
- Diverse: Cover mathematics, reasoning, coding, science
- Recent: Constructed 2023–2025, minimizing contamination risk
- Standardized: Consistent scoring (0–1 scale) across models

We **exclude** external benchmark aggregators (MMLU, GSM8K, etc.) to avoid:

1. Dataset contamination (models trained on benchmark data)
2. Saturation effects (many models achieve >90% on easier benchmarks)
3. Inconsistent evaluation protocols across research groups

B.4.11 Variable Definitions

Performance (P_{it}).

For each model i:

- *Mean*: Average score across all benchmarks where model was evaluated
- *Upper*: Maximum (best) score across benchmarks
- *Lower*: Minimum (worst) score across benchmarks

Scores are normalized to [0,1] where 1 = perfect performance.

Training Compute (C_{it}).

Total floating-point operations during pretraining:

$$C_{it} = 6 \times N \times D \tag{4}$$

where N is parameters, D is training tokens, and the factor 6 accounts for forward pass, backward pass, and parameter updates. Measured in FLOPs (floating-point operations).

Control Variables (X_{it}).

- Architecture type: Transformer, MoE, State-space, etc.
- Domain: General language, code, multimodal, etc.
- Year: 2020, 2021, ..., 2025
- Organization: OpenAI, Google, Anthropic, etc.

B.4.12 Sample Statistics

The frontier sample shows substantially higher mean performance (0.667 vs. 0.519) and larger average compute (4.82 vs. 2.45 $\times 10^{25}$ FLOPs), confirming it captures the most capable recent systems.

Our estimates align with but refine previous findings:

(Kaplan et al, 2020).

Found power law scaling with exponents around 0.05–0.10 for early GPT models. Our full sample estimate ($\beta = 0.064$) falls within this range, suggesting stable scaling relationships over 2020–2025.

Table 5: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Max
Full Sample (N=131)					
Performance (mean)	131	0.519	0.217	0.018	0.945
Training compute (10^{25} FLOPs)	131	2.45	4.12	0.001	28.5
Number of benchmarks	131	3.0	2.1	1	13
Frontier Sample (N=56)					
Performance (mean)	56	0.667	0.156	0.298	0.865
Training compute (10^{25} FLOPs)	56	4.82	5.23	0.103	28.5
Number of benchmarks	56	3.8	1.9	2	11

(Hoffmann et al, 2022).

The Chinchilla paper emphasized compute-optimal training. Our frontier estimates ($\beta = 0.159$) are higher, potentially reflecting that post-Chinchilla models better optimize the data-compute tradeoff.

(Epoch AI, 2024).

Reports continued scaling but with slowing improvements. Our heterogeneity (full vs. frontier) quantifies this precisely: historical average is $\beta = 0.064$, but leading-edge systems achieve $\beta = 0.159$.

The key contribution of our analysis is demonstrating that even the improved frontier scaling still exhibits severe diminishing returns (78× compute for 2× performance), which fundamentally constrains AI-driven growth acceleration.

B.5 Moment-Matching Calibration

The remaining four parameters—AI research efficiency ($\bar{\eta}$), critical mass threshold (M^*), AI capital formation efficiency (ξ), and AI capital depreciation (δ_M)—are jointly calibrated to match key moments characterizing frontier AI capabilities in 2024–2025.

B.5.1 Target Moments

The model is calibrated using a small set of frontier-consistent moments and normalization conditions that discipline the AI-specific parameters governing research productivity and technological progress. Given the absence of long-run macroeconomic data on artificial intelligence at the technological frontier, the calibration combines task-level empirical evidence, expert assessments, and internally consistent normalization choices, following standard practice in quantitative growth models.

Table 6 summarizes the calibration inputs. The initial AI share of research effort, $\psi_0 = 0.41$, is interpreted as a frontier-equivalent labor share capturing the effective contribution of large language models to research tasks in highly specialized domains. This value substantially exceeds economy-wide AI adoption rates, which remain in the low single digits, but is consistent with micro-level evidence documenting large productivity gains in research-relevant tasks such as coding, data analysis, and scientific writing at the frontier. Importantly, ψ_0 should not be interpreted as a measured adoption rate, but rather as an implied effective share consistent with observed task-level performance.

The initial growth rate of total factor productivity, $g_A^0 = 4.6\%$, reflects AI-augmented research productivity at the frontier. This value exceeds historical estimates of frontier TFP growth, typically in the range of 1–2%, and is intended to capture the acceleration associated with recent AI deployment rather than a directly observed macroeconomic statistic. As such, it serves as a calibration target anchoring the short-run growth implications of frontier AI systems.

Initial values for the AI capital stock, M_0, and cumulative compute stock, C_0, are normalized to the model's baseline scale. These quantities do not correspond to directly observed stocks and are chosen to facilitate identification of the remaining parameters without loss of generality.

Finally, the threshold-crossing time t^* and the long-run TFP growth rate g_A^{80} are not targeted directly but arise endogenously from the calibrated model. The reported ranges summarize the baseline implications of the calibration and provide a benchmark for the timing and magnitude of long-run growth effects under sustained AI-driven research productivity.

B.5.2 Parameter Identification

The model's structure provides clear identification of each parameter through specific moments. Base AI research efficiency $\bar{\eta}$ directly determines initial AI share ψ_0 through the relationship $\psi_0 =$

Table 6: Target Moments for Calibration (2024–2025 Frontier AI)

Moment	Target	Source
AI share of research effort (ψ_0)	0.41	Epoch AI, expert surveys
AI capital stock (M_0)	0.5	Normalized to frontier scale
Cumulative compute stock (C_0)	0.1	IEA energy data, normalized
Initial TFP growth rate (g_A^0)	0.046	OECD, AI-augmented
Threshold crossing time (t^*)	45–47 yrs	Projection models
Final TFP growth rate (g_A^{80})	0.058–0.061	Expert forecasts

Notes: Targets reflect frontier AI systems (GPT-4, Claude-3 class) as of 2025. Initial conditions M_0 and C_0 normalized to baseline scale.

$(\eta_0 M_0)/(L_A + \eta_0 M_0)$ where $\eta_0 = \bar{\eta} C_0^\beta (M_0/M^*)$. Critical mass threshold M^* controls the timing of regime transition, with higher M^* delaying threshold crossing; this parameter is directly identified by the target crossing time $t^* \approx 46$ years. AI capital formation efficiency ξ governs accumulation speed via $\dot{M} = \xi s_M Y - \delta_M M$ and is identified jointly by threshold timing and final AI capital levels. Depreciation rate δ_M affects steady-state AI capital stock and the balance between accumulation and decay, identified by final growth rates and capital-output ratios.

The parameters exhibit minimal collinearity, with $\bar{\eta}$ and M^* jointly determining initial conditions and transition timing while ξ and δ_M govern dynamic accumulation paths. Correlation matrix analysis (available upon request) confirms pairwise correlations below 0.3 for all parameter pairs.

B.5.3 Calibration Results

Simulating the model with calibrated parameters yields moments closely matching empirical targets. Table 7 reports parameter estimates with bootstrap standard errors computed via 500 replications. We draw bootstrap samples by perturbing each target moment with normally distributed noise (coefficient of variation 10%), re-calibrate for each sample, and compute standard deviations of resulting parameter distributions. The 95% confidence intervals are: $\bar{\eta} \in [0.65, 1.35]$, $M^* \in [2.0, 4.0] \times 10^6$, $\xi \in [0.30, 0.70]$, and $\delta_M \in [0.01, 0.04]$.

Table 7: Moment-Matched Parameters

Parameter	Symbol	Value	Standard Error
Base AI efficiency	$\bar{\eta}$	1.00	(0.18)
Critical mass threshold	M^*	3.0×10^6	(0.5×10^6)
AI capital formation	ξ	0.50	(0.10)
AI capital depreciation	δ_M	0.02	(0.01)

Notes: Standard errors computed via bootstrap (500 replications) assuming 10% measurement error in target moments.

The calibrated parameters generate model moments closely matching empirical targets: initial AI share is 40.9% (target: 41%), initial TFP growth is 4.6% (target: 4.6%), threshold crossing occurs at 46.0 years (target: 45–47), and final TFP growth is 5.8% (target: 5.8–6.1%). All moments fall within 5% of their targets, indicating excellent model fit.

Economic interpretation of parameter magnitudes provides face validity. Base AI efficiency $\bar{\eta} = 1.0$ implies frontier AI systems contribute research effort equivalent to one human researcher per unit of AI capital at full productivity (post-threshold). The critical mass $M^* = 3.0 \times 10^6$ represents a sixfold expansion from initial AI capital $M_0 = 0.5 \times 10^6$, achievable over 46 years under baseline investment rates. Capital formation efficiency $\xi = 0.50$ indicates that 50% of AI investment translates into productive AI capital, with the remainder lost to implementation costs, training overhead, and integration frictions. Depreciation $\delta_M = 0.02$ implies a 2% annual obsolescence rate, consistent with rapid technological turnover in AI systems where models require frequent retraining and updating.

Threshold crossing time proves relatively insensitive to β (compute scaling exponent), varying by only 1.5 years across the empirically plausible range $\beta \in [0.01, 0.15]$. Long-run growth rates, however, exhibit substantial β-sensitivity, ranging from 6.5% to 15% at the 300-year horizon. This suggests near-term predictions remain robust to scaling law uncertainty, while century-scale forecasts require continued empirical refinement of compute efficiency parameters.

B.6 External Validation: Out-of-Sample Tests

To assess the quantitative discipline of the calibrated model, we conduct a backcasting exercise. The model is calibrated using information available in 2024–2025 only, after which we backcast to 2020 and compare the resulting predictions to constructed historical benchmarks based on aggregate data, sectoral evidence, and proxy measures. In the absence of official macroeconomic time series isolating artificial intelligence, all historical values should be interpreted as approximate benchmarks rather than realized structural moments.

Table 8 reports the comparison. Trend total factor productivity growth increases from approximately 0.8% in 2020 to 1.1% in 2024, consistent with a gradual post-pandemic recovery in frontier productivity. The model predicts a 2024 growth rate of 1.0%, corresponding to a deviation of −9.1% relative to the benchmark. AI-related investment, measured as a broad macro-equivalent share of GDP encompassing software, compute infrastructure, and data capital, rises from roughly 0.5% to 2.0%. The model predicts 1.9%, implying a deviation of −5.0%.

Cumulative frontier compute, proxied by an index of effective training and deployment capacity, increases by a factor of 5.2 between 2020 and 2024. The model predicts a 4.8-fold increase, corresponding to a deviation of −7.7%. Finally, AI researcher equivalents—defined as effective labor inputs in AI-relevant research tasks rather than headcounts or workforce shares—rise sharply over the period. The model prediction closely tracks the benchmark, with a deviation of +5.0% in 2024.

Across variables, deviations remain below 10%, indicating that the calibrated model generates backcasted paths broadly consistent with available historical evidence, despite being disciplined solely by post-2024 information. This exercise should be interpreted as a consistency check rather than a formal forecast evaluation.

Table 8: Backcasting Validation Using Constructed Historical Benchmarks, 2020–2024

Variable	2020 Benchmark	2024 Benchmark	Model (2024)	Deviation
Trend TFP growth	0.8%	1.1%	1.0%	−9.1%
AI-related investment / GDP	0.5%	2.0%	1.9%	−5.0%
Cumulative frontier compute	–	5.2×	4.8×	−7.7%
AI researcher equivalents	0.1%	2.0%	2.1%	+5.0%
Mean absolute deviation	6.7%			

Notes: Benchmarks are constructed from aggregate data, sectoral evidence, and proxy measures. AI researcher equivalents denote effective research labor inputs, not workforce shares.

B.7 Parameter Uncertainty and Sensitivity

Given uncertainty in both data inputs and functional form assumptions, we assess the robustness of the model's quantitative implications through systematic sensitivity analysis. We combine Monte Carlo simulation, which propagates parameter uncertainty forward into outcome distributions, with one-at-a-time parameter variation designed to identify the parameters most influential for medium- and long-run growth outcomes.

B.7.1 Uncertainty Bands via Monte Carlo Simulation

We construct uncertainty bands for key 2050 outcomes using Monte Carlo simulation. Specifically, we draw 500 parameter vectors from normal distributions centered at baseline point estimates, with standard deviations equal to bootstrap standard errors reported in Table 7. For each draw, the model is simulated forward 25 years, generating distributions over outcomes of interest.

Table 9 reports median predictions along with the 2.5th and 97.5th percentiles of the simulated distributions. These intervals should be interpreted as conditional uncertainty bands reflecting parameter uncertainty under the maintained model structure, rather than frequentist confidence intervals.

Projected 2050 TFP growth has a median of 4.2%, with an uncertainty band spanning [3.5%, 5.4%]. Even at the lower bound, growth substantially exceeds the pre-AI historical baseline of approximately 1%. The AI share of research effort reaches a median of 77%, with an uncertainty range of [57%, 89%], indicating that AI dominance in research emerges within a generation under all but the most pessimistic parameter draws.

The knowledge stock expands by a factor of 2.8× [2.4×, 3.5×] relative to initial conditions, translating into GDP levels 120% [78%, 191%] above a counterfactual no-AI baseline. While quantitative magnitudes vary, the qualitative implication of sustained AI-driven acceleration is robust across the simulated parameter space.

Table 9: Model Predictions with 95% Confidence Intervals (2050)

Outcome	Median	Lower 95%	Upper 95%
TFP growth rate	4.20%	3.49%	5.41%
AI share of research	76.7%	57.0%	89.1%
Knowledge multiplier	2.81×	2.44×	3.49×
GDP gain vs. no-AI baseline	+120%	+78%	+191%

Notes: Confidence intervals computed via Monte Carlo (500 replications) using bootstrap standard errors from Table 7.
GDP gain measured relative to counterfactual with 1% annual growth.

Despite parameter uncertainty, the qualitative prediction of substantial AI-driven acceleration remains robust. Even the lower confidence bound implies TFP growth rates 250% above historical baselines, while the upper bound suggests potential for growth approaching 5.4% annually. The width of confidence intervals reflects genuine uncertainty about scaling parameters and AI productivity, underscoring the importance of continued empirical research on compute efficiency and AI capabilities.

Threshold crossing times are not reported in Table 9 because the critical mass $M^* = 3.0 \times 10^6$ is not reached within the 25-year horizon under baseline investment rates. In extended simulations over an 80-year horizon, the model predicts threshold crossing at a median of approximately 46 years (around 2071), with a dispersion of roughly ±8 years across parameter draws. These results are consistent with expert assessments placing transformative AI capabilities in the second half of the 21st century, though precise timing remains highly uncertain.

B.7.2 Alternative Scenarios

To further assess robustness, we examine alternative calibration scenarios that span plausible parameter ranges beyond the baseline point estimates. These scenarios are not intended to exhaust uncertainty, but rather to illustrate how the model behaves under systematically more conservative or optimistic assumptions.

The conservative scenario assumes lower scaling returns ($\beta = 0.06$ versus baseline 0.08), half the baseline AI efficiency ($\bar{\eta} = 0.5$ versus 1.0), and a higher critical mass threshold ($M^* = 5.0 \times 10^6$ versus 3.0×10^6). This configuration reflects skepticism about AI scaling laws continuing at historical rates and assumes substantial barriers to achieving autonomous AI research capabilities. The optimistic scenario takes the upper confidence interval bound for scaling ($\beta = 0.10$), 50% higher base efficiency ($\bar{\eta} = 1.5$), and a lower threshold ($M^* = 2.0 \times 10^6$), representing aggressive assumptions about AI progress and rapid capability gains.

Table 10: Scenario Comparison: 2050 Outcomes

Outcome	Conservative	Baseline	Optimistic
TFP growth rate	3.3%	4.2%	5.6%
AI share of research	47%	77%	91%
GDP gain vs. no-AI baseline	+78%	+120%	+191%
Threshold crossing (years)	68	46	31

Notes: Conservative: $\beta = 0.06$, $\bar{\eta} = 0.5$, $M^* = 5.0 \times 10^6$. Baseline: Table 7.
Optimistic: $\beta = 0.10$, $\bar{\eta} = 1.5$, $M^* = 2.0 \times 10^6$. All other parameters held at baseline.
GDP gains measured relative to counterfactual no-AI baseline with 1% annual growth.

Table 10 compares 2050 predictions across scenarios. Under conservative assumptions, TFP growth reaches 3.3% (versus 4.2% baseline), AI share reaches 47% (versus 77%), and GDP gains 78% relative to a no-AI counterfactual (versus 120%). The optimistic scenario predicts TFP growth of 5.6%, AI share of 91%, and GDP gains of 191%. Threshold crossing times range from 31 years (optimistic) to 68 years (conservative), spanning roughly four decades.

Notably, even the conservative scenario—which simultaneously assumes pessimistic values for all key parameters—predicts meaningful economic acceleration. TFP growth of 3.3% represents more than triple the pre-AI historical rate, while GDP gains of 78% within a generation would constitute a productivity acceleration comparable in magnitude to historically large episodes. This robustness across scenarios supports confidence in the model's core mechanisms: AI capital accumulation drives sustained growth acceleration, threshold crossing occurs within economically relevant timescales, and AI share eventually dominates research effort across all plausible parameter configurations.

The scenario analysis reveals that while quantitative predictions vary substantially—with 2050 GDP gains ranging from 78% to 191%—qualitative conclusions remain consistent. All scenarios predict substantial acceleration relative to historical baselines, AI accounting for a majority share of effective research inputs within this century, and threshold crossing as a discrete regime shift rather than smooth transition. This robustness suggests the model captures fundamental dynamics of AI-driven growth that persist despite irreducible uncertainty about specific parameter values.

B.7.3 One-at-a-Time Sensitivity Analysis

To identify which parameters most influence model predictions, we conduct one-at-a-time sensitivity analysis, varying each parameter ±20% while holding others fixed. Table 11 reports elasticities of 2050 TFP growth with respect to each parameter, defined as the percentage change in 2050 TFP growth divided by the percentage change in the parameter, evaluated locally around the baseline calibration.

Table 11: Parameter Sensitivity: Elasticities of 2050 TFP Growth

Parameter	Elasticity	Interpretation
ξ (capital formation)	$+0.47$	Moderate positive
ϕ (knowledge spillovers)	$+0.43$	Moderate positive
λ (research returns)	-0.40	Moderate negative
$\bar{\eta}$ (base AI efficiency)	$+0.31$	Moderate positive
M^* (critical threshold)	-0.23	Small negative
δ_M (depreciation)	-0.12	Small negative
β (compute scaling)[1]	-0.03	Negligible

Notes: Elasticities computed by varying each parameter ±20% around baseline. Positive elasticity indicates parameter increases raise TFP growth; negative elasticity indicates inverse relationship. Computed at 25-year horizon (2050).

Capital formation efficiency ξ exhibits the highest near-term elasticity at $+0.47$, indicating that a 1% increase in the rate at which AI investment translates into productive capital raises 2050 TFP growth by 0.47%. This sensitivity reflects the cumulative nature of capital accumulation: higher ξ accelerates the entire growth trajectory by enabling faster buildup of the AI capital stock. Knowledge spillovers ϕ show elasticity $+0.43$, confirming the importance of ideas building on prior knowledge in sustaining growth acceleration. The strong positive spillover effect amplifies research productivity gains, as each advance creates a larger foundation for subsequent discoveries.

The negative elasticity on research returns λ (-0.40) reflects semi-endogenous growth properties whereby stronger decreasing returns to research scale require proportionally larger research teams to maintain given growth rates, moderating acceleration. This mechanism ensures the model does not generate explosive growth: diminishing returns provide a natural brake on acceleration even as AI capabilities expand. Base AI efficiency $\bar{\eta}$ shows elasticity $+0.31$, smaller than might be expected given its direct role in determining AI productivity. This moderate sensitivity occurs because $\bar{\eta}$ affects the level of AI contribution but interacts with threshold dynamics and knowledge accumulation to determine growth rates through the scaling relationship $\eta(t) = \bar{\eta}C(t)^\beta \min(1, M/M^*)$.

The critical mass threshold M^* shows negative elasticity (-0.23): higher thresholds delay the productivity regime shift by requiring more capital accumulation before AI reaches full efficiency, reducing near-term growth. This effect operates through the $\min(1, M/M^*)$ term that scales AI efficiency below the threshold. Depreciation δ_M exhibits small negative elasticity (-0.12), as higher

[1] The compute scaling parameter β exhibits near-zero elasticity at the 25-year horizon because cumulative compute $C(25) = 0.32 < 1$, causing AI efficiency $\eta = \bar{\eta}C^\beta$ to decline slightly with higher β when $C < 1$. This counterintuitive short-run effect reverses once C exceeds unity (around year 30), after which elasticity becomes positive and grows with time. At the 80-year horizon, β elasticity reaches $+0.81$ as substantial compute accumulation ($C \approx 5.5$) magnifies scaling law effects. This horizon-dependent sensitivity reflects the time-compounding nature of compute efficiency gains.

obsolescence rates reduce steady-state AI capital through faster decay, lowering the equilibrium capital stock that supports research. The compute scaling exponent β shows negligible near-term sensitivity (-0.03) due to low cumulative compute at 25 years, though long-run sensitivity becomes substantial as discussed in the footnote.

The relatively modest elasticities—five of seven parameters below 0.5 in absolute value at the 25-year horizon—indicate that no single parameter dominates near-term outcomes. Model behavior reflects the interaction of multiple mechanisms rather than knife-edge dependence on specific calibration choices. Predictions emerge from the interaction of multiple mechanisms: capital accumulation (ξ, δ_M), knowledge production (ϕ, λ, δ), AI efficiency ($\bar{\eta}$, β), and threshold dynamics (M^*). This distributed sensitivity across structural parameters provides confidence that conclusions are not artifacts of particular calibration choices but instead reflect fundamental economic forces. The model exhibits greatest sensitivity to parameters governing capital formation and knowledge spillovers, suggesting that policies affecting investment rates and research collaboration could have first-order impacts on AI-driven growth trajectories.

B.7.4 Horizon-Dependent Sensitivity

Parameter sensitivities exhibit systematic variation across forecast horizons, reflecting the dynamic evolution of the economic system. Table 12 compares elasticities at 25-year and 80-year horizons, revealing how the relative importance of parameters shifts as AI capital accumulates and the economy transitions through the critical mass threshold.

Table 12: Parameter Sensitivity at Different Horizons

	25-Year Horizon		80-Year Horizon	
Parameter	Elasticity	Rank	Elasticity	Rank
ξ (capital formation)	$+0.47$	1	$+0.52$	2
ϕ (knowledge spillovers)	$+0.43$	2	$+0.48$	3
λ (research returns)	-0.40	3	-0.45	4
$\bar{\eta}$ (base AI efficiency)	$+0.31$	4	$+0.68$	1
M^* (critical threshold)	-0.23	5	-0.15	6
δ_M (depreciation)	-0.12	6	-0.18	5
β (compute scaling)	-0.03	7	$+0.03$	7

Notes: Elasticities computed by varying each parameter ±20% around baseline values. Rank indicates absolute magnitude of elasticity (1 = most influential). Near-term predictions show low β sensitivity as cumulative compute $C(25) = 0.32 < 1$. Long-run predictions show stronger effects as compute accumulates to $C(80) \approx 5.5$.

The most striking pattern is the dramatic rise in base AI efficiency ($\bar{\eta}$) importance, climbing from fourth rank at 25 years (elasticity $+0.31$) to first rank at 80 years (elasticity $+0.68$). This elevation reflects threshold crossing dynamics: by year 80, most parameter draws place the economy well beyond the critical mass M^*, activating the full productivity regime where $\eta = \bar{\eta}C^{\beta}$ without the dampening factor M/M^*. Once the threshold is crossed, base efficiency directly determines AI research productivity without attenuation, magnifying its influence on long-run growth. Capital formation efficiency ξ maintains top ranking at both horizons (elasticity rising modestly from $+0.47$ to $+0.52$), confirming its persistent importance throughout the transition. The cumulative nature of capital accumulation ensures that ξ affects the entire growth trajectory, making it the most consistently influential parameter across time scales.

Knowledge spillovers ϕ and research returns λ show moderate elasticity increases (from $+0.43$ to $+0.48$ and from -0.40 to -0.45 respectively), indicating their effects strengthen gradually as the knowledge stock expands and research teams grow. These parameters govern the ideas production function, and their influence compounds as larger knowledge bases amplify spillover effects while larger research teams encounter steeper diminishing returns. The critical mass threshold M^* exhibits declining importance in absolute terms (elasticity magnitude falling from 0.23 to 0.15), a counterintuitive result explained by timing effects: at 25 years, many parameter draws place the economy near the threshold where small changes in M^* significantly affect whether crossing has occurred, while at 80 years most draws are well past the threshold regardless of M^* value, reducing sensitivity.

The compute scaling parameter β displays the most dramatic horizon dependence, shifting from slightly negative elasticity (-0.03) at 25 years to slightly positive ($+0.03$) at 80 years, though remaining quantitatively small at both horizons. This non-monotonic pattern arises from the mathematical properties of the scaling relationship $\eta = \bar{\eta}C^\beta$. At year 25, cumulative compute $C = 0.32 < 1$, causing the derivative $\partial C^\beta/\partial\beta = C^\beta \ln(C)$ to be negative (since $\ln(C) < 0$ when $C < 1$). Higher β thus reduces AI efficiency in the short run, creating negative sensitivity. Around year 30, C crosses unity and the effect reverses sign, but elasticity remains near zero until substantial compute accumulation occurs. By year 80, $C \approx 5.5$ and elasticity turns positive, though the logarithmic growth of C^β keeps the effect modest. Extended simulations to year 200 (where $C \approx 50$) show β elasticity rising to $+0.81$, confirming that compute scaling effects compound exponentially over very long horizons but contribute little to near-term and medium-term dynamics.

Depreciation δ_M shows modestly increasing absolute influence (from -0.12 to -0.18), reflecting the growing importance of capital stock maintenance as AI capital levels rise. At 25 years when $M \approx 1.2$, depreciation flows are small relative to investment flows, limiting sensitivity. By year 80 when $M \approx 20$, depreciation represents a substantial drain on capital accumulation ($\delta_M M \approx 0.4$ per year), making the parameter more consequential. This pattern illustrates how stock-flow dynamics shift in importance as the economy matures: in early stages, accumulation dominates and flow parameters matter most, while in later stages, stock maintenance becomes critical and stock-related parameters gain influence.

The horizon-dependent rankings reveal that optimal research and policy priorities shift across time scales. Near-term acceleration depends primarily on capital formation efficiency (ξ) and knowledge spillover strength (ϕ), suggesting that policies promoting AI investment and research collaboration deliver immediate benefits. Long-run growth increasingly hinges on base AI efficiency ($\bar{\eta}$), highlighting the importance of fundamental research on AI capabilities and algorithmic improvements that raise the productivity ceiling. The diminishing importance of the threshold parameter M^* at long horizons suggests that uncertainty about the exact threshold level—while crucial for timing predictions—matters less for ultimate growth outcomes than uncertainty about efficiency and spillover parameters. This finding provides reassurance that long-run forecasts are robust to threshold calibration despite near-term sensitivity.

B.8 Model Validation Against Stylized Facts

Before proceeding to policy analysis, we assess whether the calibrated model reproduces a set of well-documented stylized facts from the macroeconomic growth and technological change literature. This exercise provides external validation by evaluating whether the model's structural mechanisms generate empirical regularities that were not directly targeted in calibration.

First, research and development effort has increased dramatically over the postwar period—by roughly an order of magnitude since 1950—while long-run aggregate growth rates have remained broadly stable (Jones, 1995; Bloom et al, 2020). The model captures this pattern through its semi-endogenous growth structure, with $\phi = 0.52 < 1$, which implies that long-run growth depends on population growth rather than the scale of research effort. In the absence of AI augmentation, the model generates approximately constant TFP growth of around 1% annually despite rising R&D employment, consistent with historical experience and validating the specification of the ideas production function.

Second, measured TFP growth has slowed over recent decades, declining from postwar rates near 2% to values below 1% in the 2000s and 2010s (Gordon, 2016). This pattern is consistent with the "ideas getting harder to find" hypothesis (Bloom et al, 2020), which the model embeds through diminishing returns to research effort ($\lambda < 1$). As research scale expands, increasingly large inputs are required to sustain a given rate of technological progress. Under the baseline calibration, the model produces pre-AI TFP growth rates near 0.8% by 2020 and modest acceleration thereafter with the introduction of AI-driven research productivity, consistent with recent trend estimates in the productivity literature.

Third, labor's share of national income has declined in many advanced economies since the early 1980s, falling by roughly 5–7 percentage points in the United States (Karabarbounis and Neiman, 2014). By modeling AI as a form of capital-augmenting technological change operating through the research channel, the model endogenously generates a declining labor share as AI capital accumulates. Under the calibrated parameters, the model produces labor share dynamics that closely track the magnitude and timing of the observed decline, supporting the interpretation of AI and automation as disproportionately benefiting capital relative to labor.

Fourth, computational resources devoted to training frontier AI systems have grown extremely rapidly over the past decade, with successive generations of state-of-the-art models requiring orders of magnitude more compute (Epoch AI, 2025). The model reproduces this qualitative pattern through the interaction of rising AI investment shares and declining effective compute costs. Under the baseline calibration, simulated compute accumulation exhibits sustained exponential growth over the 2010–2025 period at rates comparable to those documented in frontier training datasets, lending support to the model's compute accumulation mechanism.

Fifth, income inequality has increased in most advanced economies since 1980, as reflected in rising top

B.9 Limitations and Scope Conditions

Despite its empirical discipline, the analysis is subject to several important limitations that delimit the scope of the results and highlight priorities for future research.

Data scarcity and measurement constraints.

The calibration relies primarily on information from the 2023–2025 period, during which large-scale frontier AI systems have only recently begun to operate in research-intensive settings. While state-of-the-art language models provide clear evidence of substantial AI research capabilities, systematic measurement of their productivity contributions remains limited. Available evidence is drawn from a combination of expert assessments, task-level studies, and sectoral indicators, which are informative but lack the statistical depth of traditional macroeconomic datasets. As AI deployment becomes more widespread and standardized productivity measures emerge, future recalibration using richer data could sharpen parameter estimates and reduce uncertainty.

Scaling law uncertainty and extrapolation risk.

The compute scaling parameter $\beta = 0.08$ is estimated using data from models trained on compute levels up to approximately 10^{25} FLOPs. Long-horizon projections implicitly extrapolate these relationships to substantially higher compute regimes. Physical constraints, algorithmic saturation, data bottlenecks, or energy limitations could weaken scaling at extreme levels, effects that are not captured by current empirical estimates. Because long-run growth outcomes are particularly sensitive to β, continued empirical validation of scaling relationships remains critical. The scenario analysis spanning $\beta \in [0.06, 0.10]$ mitigates this concern by bracketing plausible outcomes, but uncertainty regarding extreme-scale extrapolation cannot be fully resolved with existing evidence.

Abstraction from institutional and adoption frictions.

The model abstracts from regulatory, organizational, and political economy frictions that may slow or redirect AI deployment. AI capital accumulation is modeled as a smooth function of investment and depreciation, whereas real-world diffusion may be constrained by safety regulation, labor market rigidities, firm-level adjustment costs, or social resistance in sensitive sectors. As a result, the model's projections should be interpreted as conditional on relatively supportive institutional environments. Incorporating explicit adoption frictions calibrated to historical technology diffusion patterns would provide a more detailed characterization of transition dynamics.

Threshold identification and regime uncertainty.

The critical mass threshold $M^* = 3.0 \times 10^6$ is calibrated to align model dynamics with plausible transition timing rather than identified from direct observations of discrete capability jumps. While theoretical and expert-based arguments suggest the possibility of nonlinear regime changes in AI capabilities, both the existence and sharpness of such thresholds remain uncertain. The model adopts a stylized threshold structure for tractability. Sensitivity analysis indicates that threshold timing varies moderately across parameter draws, but uncertainty regarding the nature of AI capability transitions remains an inherent limitation.

Exogenous treatment of AI investment dynamics.

In the baseline specification, the AI investment share s_M follows an exogenously specified path based on recent trends. In practice, investment decisions are endogenous and respond to realized productivity gains, expectations, and financing conditions. Endogenizing s_M within an optimal investment or savings framework could capture feedback effects that amplify or dampen AI-driven growth. The current approach abstracts from these mechanisms and therefore yields conservative projections by ruling out investment acceleration in response to AI success.

Taken together, these limitations define the scope of the analysis rather than undermining its core conclusions. Within these bounds, the model provides a coherent and empirically disciplined framework for studying AI-driven growth. The results indicate that observed AI capabilities and recent economic trends are consistent with substantial future growth acceleration under a wide range of plausible assumptions. While uncertainty remains—particularly regarding long-run scaling and institutional responses—the central mechanism linking AI capital accumulation to research productivity gains is well supported by available evidence.

The relation between the model's growth mechanism and alternative automation-based interpretations is discussed separately in Appendix C.

C Relation to Alternative Growth Mechanisms

This section clarifies how the mechanisms emphasized in the model relate to alternative views of technological change and economic growth, particularly those emphasizing automation, task substitution, and historical stability of aggregate productivity growth. The purpose is not to adjudicate among competing theories, but to situate the model's assumptions and implications relative to well-established perspectives in the growth literature.

Semi-endogenous growth and bounded long-run dynamics.

The model is explicitly semi-endogenous. In the absence of artificial intelligence, long-run growth depends on population growth and exhibits a constant balanced growth path despite rising research effort. This property follows from diminishing returns to research scale ($\lambda < 1$) and imperfect knowledge spillovers ($\phi < 1$), consistent with standard results in the semi-endogenous growth literature. The introduction of AI does not alter this fundamental structure: it generates transitional dynamics and level effects, but does not mechanically produce unbounded or permanently exponential growth.

Endogenous emergence of regime transitions.

Nonlinear dynamics arise from the interaction between AI capital accumulation, compute scaling, and the ideas production process. The model's regime transition is not imposed exogenously, but emerges analytically when accumulated AI capital reaches a scale at which its effective contribution to research ceases to be constrained. Calibration pins down the timing of this transition using observed AI investment flows and frontier research contributions. Sensitivity analysis shows that while the timing of threshold crossing varies across parameter configurations, the qualitative presence of nonlinear transition dynamics does not depend on knife-edge assumptions.

Compatibility with short-run productivity skepticism.

The model accommodates the empirical observation that aggregate productivity growth remained modest through 2024 despite rapid advances in AI capabilities. In the calibrated economy, AI initially operates below full efficiency due to limited capital stock and cumulative compute. As a result, early deployment generates small measured aggregate effects, consistent with recent macroeconomic evidence. Larger effects arise only gradually as AI capital accumulates and interacts with the research sector, reconciling micro-level productivity gains with muted short-run macro outcomes.

Conservative treatment of compute scaling.

Baseline simulations rely on conservative estimates of compute scaling elasticities derived from the full sample of post-2020 models. Stronger scaling observed among frontier systems is documented but not adopted as a baseline input. Moreover, near- and medium-run dynamics are shown to be largely insensitive to the scaling parameter because cumulative compute remains below unity for several decades. This limits the role of extrapolation from frontier performance trends in shaping near-term quantitative results.

Growth without reliance on labor displacement.

Growth acceleration in the model operates primarily through the research channel rather than through direct substitution of labor in production. The model does not require large contemporaneous job displacement or rapid task automation to generate higher growth rates. This distinguishes it from task-based automation models in which productivity gains hinge on widespread labor replacement, and aligns the mechanism more closely with historical episodes of research-driven technological change.

Historical stability as an explained outcome.

The model reproduces decades of stable aggregate growth, rising research effort, and declining research productivity by construction. AI alters this trajectory only when its accumulated capital and compute reach economically meaningful scale. In this sense, the framework explains why growth was stable for an extended period and why it may change, rather than assuming instability as a primitive feature of the economy.

Conditional interpretation of quantitative results.

All quantitative implications are conditional on sustained investment in AI, continued deployment in research-intensive domains, and the absence of binding institutional or regulatory constraints. The results should therefore be interpreted as conditional trajectories under supportive environments rather than unconditional forecasts. This framing aligns the model with conservative interpretations of technological change that emphasize institutional mediation rather than technological determinism.

Taken together, these features ensure that the model remains consistent with conservative views on long-run growth while allowing for the possibility that frontier AI alters transitional dynamics through well-defined economic mechanisms. The contribution is not to assert that AI must transform growth, but to show that—given observed frontier capabilities and conservative assumptions—substantial effects are internally consistent, empirically disciplined, and theoretically coherent.

C.1 Conclusion

This calibration exercise shows that a tractable growth framework can be disciplined using frontier AI evidence while remaining consistent with established macroeconomic dynamics. The model matches a small set of calibration targets characterizing AI capabilities and economic conditions in 2024–2025 and reproduces a range of well-documented empirical regularities that were not explicitly targeted. Parameter estimates are economically interpretable and stable across alternative identification approaches. In particular, the calibrated base AI efficiency $\bar{\eta} = 1.0$ implies that frontier AI capital contributes research effort comparable to a human researcher at full efficiency, while the critical mass $M^* = 3.0 \times 10^6$ corresponds to a scale expansion that is achievable within several decades under baseline investment paths.

Under the baseline calibration, the model predicts a marked acceleration in productivity growth over the coming decades. TFP growth reaches 4.2% by 2050, compared to a pre-AI trend near 1%, while the AI share of research effort rises to approximately 77%. Monte Carlo simulations that propagate parameter uncertainty yield a 2050 TFP growth range of [3.5%, 5.4%], indicating that substantial acceleration arises across a wide set of plausible parameter configurations. Scenario analysis further shows that even conservative assumptions generate large macroeconomic effects, with projected 2050 GDP levels 78% above a no-AI counterfactual, while more optimistic scenarios imply gains approaching 191%.

Sensitivity analysis highlights that model predictions are not driven by any single parameter. Near-term outcomes depend most strongly on capital formation efficiency ξ and knowledge spillovers ϕ, whereas long-run growth becomes increasingly sensitive to base AI efficiency $\bar{\eta}$ once the economy moves beyond the critical mass threshold. This horizon dependence underscores that policy relevance is time-varying: institutional arrangements that support investment and research collaboration matter most in the near term, while advances in fundamental AI capabilities play a larger role in shaping long-run growth trajectories.

The model's ability to align with multiple empirical patterns while maintaining theoretical coherence suggests it provides a useful framework for evaluating the macroeconomic implications of AI. While uncertainty remains—particularly regarding long-run scaling dynamics and institutional responses—the calibrated structure offers a disciplined basis for policy analysis, which we turn to next.

References

Anthropic (2025) Ai economic impact index. https://www.anthropic.com, accessed January 2025

Bloom N, Jones CI, Van Reenen J, et al (2020) Are ideas getting harder to find? American Economic Review 110(4):1104–1144

Bureau of Labor Statistics (2025) Labor productivity and costs. https://www.bls.gov

Epoch AI (2024) Trends in training compute of machine learning systems. Tech. rep., Epoch AI, URL https://epochai.org/trends

Epoch AI (2025) Growth, automation, and training efficiency (gate) model. Tech. rep., Epoch AI, URL https://epochai.org/gate

Gollin D (2002) Getting income shares right. Journal of Political Economy 110(2):458–474

Gordon RJ (2016) The Rise and Fall of American Growth. Princeton University Press

Hoffmann J, Borgeaud S, Mensch A, et al (2022) Training compute-optimal large language models. arXiv preprint ArXiv:2203.15556

IEA (2025) Energy technology perspectives 2025. https://www.iea.org

Jones CI (1995) R&d-based models of economic growth. Journal of Political Economy 103(4):759–784

Jones CI (2002) Sources of u.s. economic growth in a world of ideas. American Economic Review 92(1):220–239

Kaplan J, McCandlish S, Henighan T, et al (2020) Scaling laws for neural language models. arXiv preprint ArXiv:2001.08361

Karabarbounis L, Neiman B (2014) The global decline of the labor share. Quarterly Journal of Economics 129(1):61–103

MLCommons (2025) MLPerf benchmark results. https://mlcommons.org, accessed January 2025

OECD (2025) Main science and technology indicators. https://www.oecd.org/sti/msti.htm

Solow RM (1956) A contribution to the theory of economic growth. Quarterly Journal of Economics 70(1):65–94

Stanford HAI (2025) Artificial intelligence index report 2025. https://aiindex.stanford.edu

World Bank (2025) World development indicators. https://data.worldbank.org, accessed January 2025