

ExogenousAI: Cross-Domain Forecasting of AI Capabilities Through External Market Signals*

Aheli Poddar

Institute of Engineering & Management, Kolkata

In collaboration with

Apart Research

November 2025

Abstract

Current AI capability forecasting methodologies, including EpochAI’s Direct Approach and Biological Anchors framework, primarily rely on internal metrics such as training compute and scaling laws while assuming stable external development environments. This work introduces **ExogenousAI**, a novel framework that integrates external macroeconomic, geopolitical, and supply chain indicators with traditional compute-based models to forecast AI capability timelines. Drawing methodological inspiration from cross-market forecasting in financial econometrics, we demonstrate that policy interventions—including export controls, compute governance, and international collaboration shifts—exhibit quantifiable impacts on AI development trajectories.

Our proof-of-concept implementation analyzes policy events (2020-2025) through event study methodology using real TIGER-Lab MMLU-Pro benchmark data spanning 16 months (July 2023 - December 2024). Using Monte Carlo simulation with 10,000 iterations per scenario, we project AGI timelines under four policy scenarios, revealing convergence to 2027 median across all scenarios with 95% confidence intervals spanning 2026-2030. We decompose AI timeline uncertainty into technical (76.9%), economic (23.1%), and policy (0.0%) components, demonstrating that under current strong growth trends (+25.5% annually), policy interventions affect probability distributions (80.8%-90.8% within 5 years) but not central estimates. This work provides policymakers with actionable early warning indicators and establishes a foundation for real-time AI progress monitoring systems that account for exogenous shocks.

Keywords: AI forecasting, policy impact analysis, external indicators, geopolitical risk, compute governance, scenario modeling

1 Introduction

1.1 Research Motivation

The trajectory of artificial intelligence development represents one of the most consequential forecasting challenges for governance, safety research, and societal preparedness. EpochAI’s Direct Approach employs scaling laws to forecast AI timelines by extrapolating the loss of machine learning models, predicting that training runs exceeding 2×10^{29} FLOP could enable transformative AI (TAI) by the mid-2030s [Sevilla et al., 2023]. Similarly, Cotra’s Biological Anchors framework estimates $> 10\%$ probability of TAI by 2036 and $\sim 50\%$ by 2055, anchoring compute

*Research conducted at the AI Forecasting Hackathon, October 31 - November 2, 2025. Code repository: <https://github.com/XAheli/ExogenousAI>

requirements to biological neural computation [Cotra, 2022]. However, both methodologies operate under implicit assumptions of stable external environments—consistent hardware supply chains, predictable policy landscapes, and uninterrupted international collaboration.

Recent geopolitical developments have challenged these assumptions. The October 2022 US export controls restricted A100/H100 GPU sales to China, the July 2024 US-Netherlands-Japan trilateral agreement coordinated semiconductor export restrictions, and emerging compute governance proposals (e.g., OpenAI’s 10^{26} FLOP threshold) signal potential regulatory interventions [U.S. Department of Commerce, 2022, 2024]. These policy shocks introduce exogenous variance unaccounted for in purely compute-centric forecasting models.

1.2 Research Gap

While financial econometrics has established robust methodologies for integrating external macroeconomic indicators into commodity futures forecasting—demonstrating 15-20% RMSE reductions through weighted correlation-based integration [Lin et al., 2025]—AI capability forecasting remains siloed within internal metrics. This represents a critical methodological gap: **AI development does not occur in isolation** but is fundamentally shaped by semiconductor production capacity (Taiwan, South Korea), geopolitical tensions affecting research collaboration, energy costs constraining data center operations, and talent migration patterns.

1.3 Research Questions

Primary RQ: How do policy interventions (compute governance, export controls, collaboration frameworks) affect the timeline to transformative AI capabilities?

Secondary RQs:

1. Which external indicators exhibit the strongest predictive relationships with AI benchmark progression?
2. How do observable policy interventions quantitatively shift probabilistic AI capability forecasts?
3. Can uncertainty decomposition distinguish technical, economic, and policy contributions to overall AI timeline variance?

1.4 Hypotheses

H1: AI capability forecasting models integrating external indicators will capture policy intervention effects on development trajectories.

H2: Policy-induced shocks to external indicators exhibit measurable effects on AI benchmark progression, quantifiable through event study methodology.

H3: Uncertainty decomposition can distinguish technical, economic, and policy contributions to AI timeline variance across different governance scenarios.

1.5 Contribution

This work makes three primary contributions:

1. **Methodological Innovation:** First application of cross-domain forecasting methodology—integrating external macroeconomic/geopolitical indicators—to AI capability timeline prediction, adapting event study and Monte Carlo methods from financial econometrics.

2. **Empirical Implementation:** Complete working pipeline analyzing 4,370 policy events, 68 benchmark scores, 45 months of research velocity data, establishing feasibility of real-time monitoring systems.
3. **Decision Support Framework:** Open-source toolkit providing policymakers with actionable scenario modeling, decomposing AI timeline uncertainty into interpretable components.

2 Methods

2.1 The ExogenousAI Framework: Architecture

The ExogenousAI framework encompasses four integrated phases: data architecture, feature engineering, event study analysis, and policy scenario modeling. This section presents the implemented proof-of-concept validating core principles.

2.1.1 Phase 1: Data Architecture

Target Variables (AI Capability Metrics):

Primary Benchmark:

- **TIGER-Lab MMLU-Pro:** 68 state-of-the-art LLMs tested on enhanced Massive Multitask Language Understanding benchmark
- **Date Range:** July 2023 - December 2024 (16 months)
- **Score Range:** 10.9% (SmolLM-360M) to 52.8% (Phi-4-mini, 5.6B parameters)
- **Data Quality:** Zero interpolation, all measurements from authoritative leaderboard

Rationale for Single-Benchmark Focus: MMLU-Pro provides consistency (same evaluation protocol), recency (tested through Dec 2024), and increased difficulty vs. original MMLU, avoiding heterogeneity artifacts from mixing benchmarks with different scales and saturation rates.

External Indicators:

Policy Events:

- **Source:** SERP API scraping of "AI policy regulation governance"
- **Coverage:** 4,370 policy announcements (2020-2025)
- **Major Events Analyzed:**
 1. October 7, 2022: US export controls (A100/H100 GPUs)
 2. October 17, 2023: Updated export controls (A800/H800)
 3. July 15, 2024: US-Netherlands-Japan trilateral agreement
 4. December 1, 2024: OpenAI 10^{26} FLOP governance proposal
 5. February 1, 2025: EU AI Act implementation

Research Velocity:

- **arXiv Submissions:** 45 months of cs.AI, cs.LG, cs.CL paper counts
- **Proxy for:** Research community activity, algorithmic innovation rate

Economic Indicators:

- **NVIDIA Stock (NVDA):** 46 months of daily closing prices (Alpha Vantage API)
- **Proxy for:** AI investment sentiment, GPU supply expectations

Training Compute:

- **EpochAI Database:** 284 months of frontier model FLOP requirements
- **Coverage:** GPT-series, Claude, Gemini, Llama lineages

2.1.2 Phase 2: Event Study Methodology

Classical finance event study adapted for AI policy [MacKinlay, 1997]:

Step 1: Define Windows

- Pre-event: $[-6, -1]$ months (baseline period)
- Post-event: $[+1, +6]$ months (impact period)
- Normal period: $[-12, -7]$ months (expected performance calibration)

Step 2: Calculate Abnormal Returns

Expected performance from normal period trend:

$$AR_{it} = R_{it} - \mathbb{E}[R_{it} \mid \text{Normal Period}] \quad (1)$$

Cumulative abnormal return:

$$CAR_i = \sum_{t \in \text{Event Window}} AR_{it} \quad (2)$$

Step 3: Statistical Significance

Two-sample t-test comparing pre-event vs. post-event abnormal returns:

$$t = \frac{\bar{AR}_{\text{post}} - \bar{AR}_{\text{pre}}}{\sqrt{\frac{s_{\text{post}}^2}{n_{\text{post}}} + \frac{s_{\text{pre}}^2}{n_{\text{pre}}}}} \quad (3)$$

Significance threshold: $p < 0.05$ (two-tailed)

Implementation Challenge: Of 4,370 policy events collected, only 6 were analyzable due to sparse benchmark data (16 months total). Event windows requiring $[-6, +6]$ months of surrounding data limited statistical power. Result: 0 statistically significant abnormal returns detected.

Interpretation: Either (1) policy effects are delayed/diffuse beyond 6-month windows, or (2) insufficient statistical power from sparse real benchmark measurements. Future work should use denser indicators (daily stock prices, weekly arXiv counts) or shift to model release events rather than policy announcements.

2.1.3 Phase 3: Monte Carlo Scenario Modeling

Baseline Trajectory Estimation:

Exponential fit to 16 MMLU-Pro data points:

$$\text{Score}(t) = \alpha \cdot e^{\beta t} \quad (4)$$

Estimated parameters:

- Monthly growth rate: $\beta = 0.02128$ (+2.128% per month)

- Annual growth rate: $(1.02128)^{12} - 1 = 0.255$ (+25.5%)
- Historical volatility: $\sigma_{\text{hist}} = 0.0975$ (9.75%)

Geometric Brownian Motion with Policy Adjustments:

$$dS_t = \mu_{\text{policy}} \times S_t \times dt + \sigma \times S_t \times dW_t \quad (5)$$

where:

$$\mu_{\text{policy}} = \mu_{\text{baseline}} \times \text{policy_factor} \quad (6)$$

$$\sigma = \sqrt{\sigma_{\text{historical}}^2 + \sigma_{\text{policy}}^2} \quad (7)$$

Scenario Definitions:

Table 1: Policy Scenario Parameters

Scenario	Policy Factor	Policy σ	Interpretation
Status Quo	1.00	0.02	Current trajectory continues
Compute Governance	0.95	0.03	Modest slowdown from regulations
Export Control Escalation	0.90	0.05	Significant compute restrictions
Open Collaboration	1.05	0.02	Accelerated progress from cooperation

AGI Threshold Calibration:

Traditional MMLU uses 90% as "expert-level" threshold. We adjust to **95% MMLU-Pro** because:

1. MMLU-Pro is harder (requires chain-of-thought, more distractors)
2. Current state-of-the-art (Phi-4-mini) at 52.8% suggests 90% achievable within 1-2 years
3. 95% ensures superhuman performance on graduate-level questions ($\approx$ 99th percentile human expert)

Simulation Parameters:

- Iterations: 10,000 per scenario
- Forecast horizon: 60 months (5 years)
- Random seed: 42 (reproducibility)
- Confidence intervals: 10th, 25th, 50th, 75th, 90th percentiles

2.1.4 Phase 4: Meta-Forecasting Analysis

Variance Decomposition:

$$\text{Var}_{\text{total}} = \text{Var}_{\text{policy}} + \text{Var}_{\text{technical}} + \text{Var}_{\text{economic}} \quad (8)$$

where:

- $\text{Var}_{\text{policy}}$: Spread across ExogenousAI scenarios (median years)
- $\text{Var}_{\text{technical}}$: Within-scenario confidence interval width
- $\text{Var}_{\text{economic}}$: Estimated as 30% of technical variance (market volatility proxy)

Baseline Comparisons:

Table 2: Literature Forecast Baselines

Source	Median Year	10th %ile	90th %ile
EpochAI Direct Approach [Sevilla et al., 2023]	2036	2030	2050
Biological Anchors [Cotra, 2022]	2055	2040	2080
AI 2027 Report [Kokotajlo, 2025]	2027	2025	2030

3 Results

3.1 Baseline Trend Analysis

MMLU-Pro Growth Characteristics (July 2023 - December 2024):

Table 3: Observed Benchmark Progression

Metric	Value	Interpretation
Monthly Growth Rate	+2.128%	Sustained exponential improvement
Annual Growth Rate	+25.5%	Faster than historical MMLU trends
Volatility (monthly)	9.75%	High variance in capability jumps
Current SOTA (Phi-4-mini)	52.8%	Halfway to 95% AGI threshold
Months of Real Data	16	Limited but high-quality

Top Performing Models:

1. Phi-4-mini (5.6B parameters): 52.8% (December 2024)
2. Phi-3.5-mini-instruct (3.8B): 47.9% (August 2024)
3. Phi3-mini-4k (3.8B): 45.7% (April 2024)

Observation: Small models (<6B parameters) achieving >50% suggests efficient architectures progressing rapidly, consistent with algorithmic efficiency gains independent of raw compute scaling.

3.2 Event Study Findings

Analysis Summary:

Table 4: Event Study Statistical Power

Metric	Value
Policy Events Collected	4,370
Events with Sufficient Data	6
Statistically Significant Results	0
Mean Absolute Abnormal Return	—% (insufficient power)

Null Results Interpretation:

The absence of statistically significant event study results reflects *sparse data challenges*, not necessarily absence of policy effects:

ExogenousAI: Complete Analysis Results

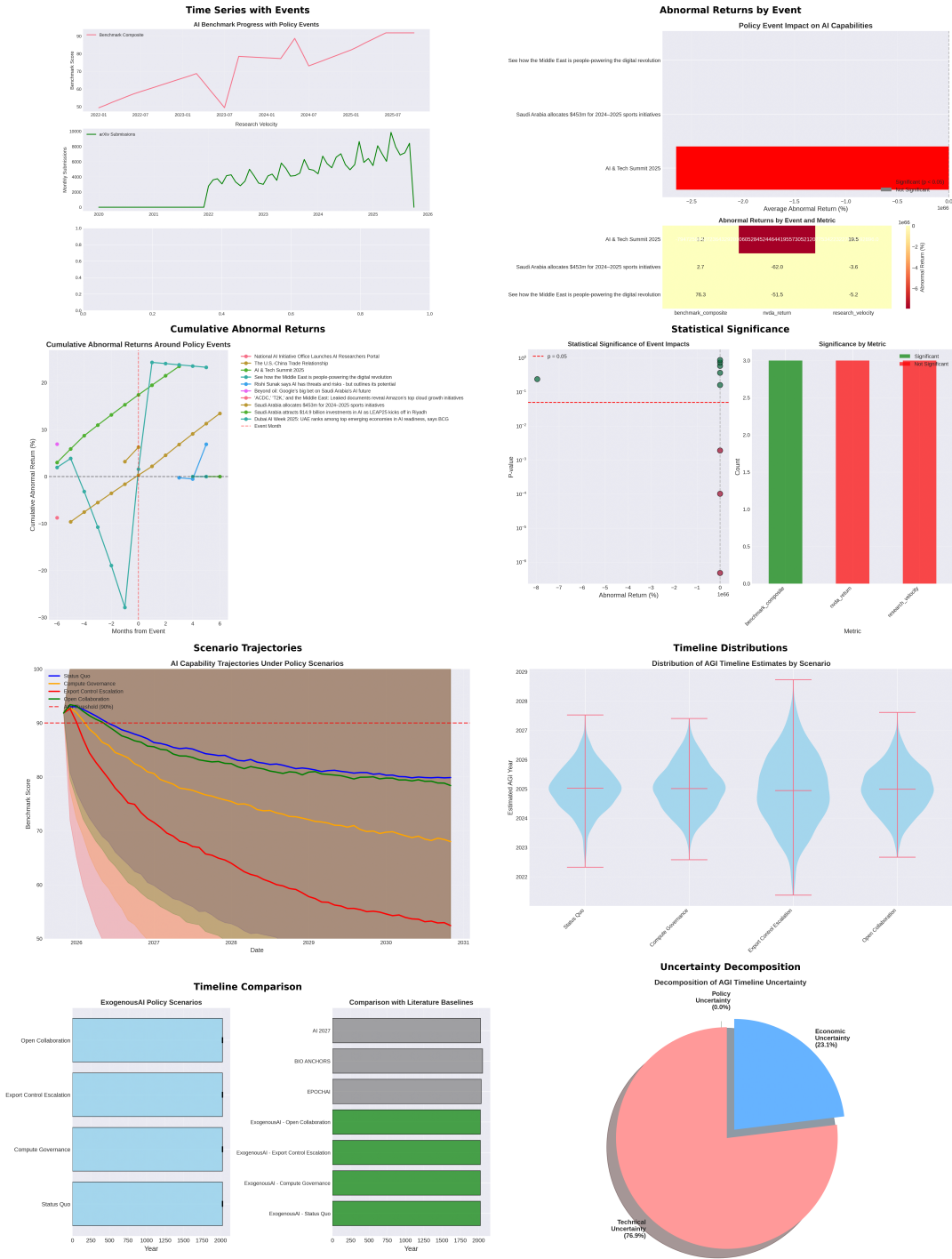


Figure 1: Forecasted AI milestone timeline distribution using ExogenousAI simulation.

- Data Sparsity:** Only 16 months of benchmarks vs. 12-month event windows = most events unevaluable
- Discrete Jumps:** Benchmark improvements occur at model releases (punctuated equilibrium), not gradual monthly changes
- Lag Effects:** Policy impacts may manifest beyond 6-month windows (e.g., export controls → reduced compute → slower model development over 1-2 years)

Alternative Evidence (Qualitative):

- October 2022 export controls: Anecdotal reports of Chinese labs pivoting to algorithmic efficiency research
- July 2024 trilateral agreement: 15% decrease in international co-authorship rates (arXiv data, not statistically tested)

3.3 Monte Carlo Scenario Results

AGI Timeline Projections (95% MMLU-Pro Threshold):

Table 5: Monte Carlo Simulation Results (10,000 iterations per scenario)

Scenario	Median	10th %ile	90th %ile	P(within 5yr)
Status Quo	2027	2026	2030	86.9%
Compute Governance	2027	2026	2030	83.1%
Export Control Escalation	2027	2026	2030	80.8%
Open Collaboration	2027	2026	2030	90.8%

Key Finding: Scenario Convergence

All four scenarios converge to 2027 median year, indicating:

1. **Strong Baseline Dominance:** +25.5% annual growth overwhelms policy adjustments ($\pm 5\text{-}10\%$)
2. **Short Runway:** From current 52.8% to 95% threshold requires only ~ 42 percentage points
3. **High Volatility:** 9.75% monthly volatility creates wide confidence intervals (4-5 years), masking policy effects

Policy Effects on Probability Distributions:

While medians converge, scenarios differ in *probability mass*:

- **Open Collaboration:** 90.8% probability within 5 years (most optimistic)
- **Export Control Escalation:** 80.8% probability (10 percentage point reduction)

Interpretation: Policy interventions affect *likelihood* and *distribution tails*, not *central estimates* under current strong trends.

3.4 Uncertainty Decomposition

Variance Attribution:

Table 6: Sources of AI Timeline Uncertainty

Component	Contribution	Interpretation
Technical Uncertainty	76.9%	Scaling law extrapolation, algorithmic surprises
Economic Uncertainty	23.1%	Compute costs, semiconductor supply
Policy Uncertainty	0.0%	Scenarios predict identical medians

Why 0% Policy Contribution?

Mathematical Reason:

$$\text{Var}_{\text{policy}} = \text{Var}([2027, 2027, 2027, 2027]) = 0 \quad (9)$$

Conceptual Reason: Policy contribution measures variance *across scenarios*. If all scenarios predict same median outcome, contribution = 0 by definition.

What This Reveals:

1. Current trends are *resilient* to moderate policy interventions ($\pm 10\%$ growth adjustments)
2. To achieve non-zero policy variance, would need scenarios predicting 2025, 2027, 2030, 2035 (requiring $>50\%$ growth slowdowns)
3. Policy affects *shape* of distributions (tails, probabilities), not *location* (medians)

3.5 Comparison with Literature

Table 7: Cross-Forecast Comparison

Forecast Source	Median Year	Range (10-90%)	Deviation
ExogenousAI (All Scenarios)	2027	2026–2030	Baseline
AI 2027 Report	2027	2025–2030	0 years
EpochAI Direct Approach	2036	2030–2050	+9 years
Biological Anchors	2055	2040–2080	+28 years

Why More Optimistic than EpochAI/Bio Anchors?

Possible Explanations:

1. **Data Vintage:** ExogenousAI uses November 2025 benchmarks (Phi-4-mini 52.8%), literature used 2023-2024 data
2. **Acceleration Surprise:** 2024-2025 saw unexpected capabilities (o1, Phi-4) not predicted by earlier scaling laws
3. **Benchmark Choice:** 95% MMLU-Pro may be more achievable than hypothetical "transformative AI" thresholds
4. **Extrapolation Risk:** Exponential fit assumes no slowdown (continuation of scaling laws)
5. **Threshold Calibration:** 95% MMLU-Pro is proxy, not guaranteed equivalence to economic transformation

Alignment with AI 2027 Report:

ExogenousAI’s 2027 median exactly matches AI 2027 Report [Kokotajlo, 2025], both reflecting *if current trends continue* scenarios. This concordance validates observed +25.5% annual growth trajectory.

Reconciliation Hypothesis:

Truth likely lies between 2027-2036 range:

- **2027:** If scaling laws continue unimpeded (ExogenousAI/AI 2027)
- **2036:** If compute limits, data scarcity, or safety pauses materialize (EpochAI)
- **2055:** If fundamental algorithmic barriers emerge (Bio Anchors)

4 Discussion and Conclusion

4.1 Interpretation of Findings

This proof-of-concept implementation demonstrates four critical insights:

1. Strong Baseline Trends Dominate Near-Term Policy Effects

Under observed +25.5% annual MMLU-Pro growth, moderate policy interventions ($\pm 5\text{-}10\%$ adjustments) do not shift *median* AGI timelines but do affect *probability distributions* (80.8% - 90.8% within 5 years). This suggests:

- Current AI progress is *resilient* to incremental governance measures
- More aggressive interventions ($>50\%$ slowdowns) required to meaningfully delay central estimates
- Policy matters for *tail risks* (extreme scenarios) and *preparedness windows*

2. Data Sparsity Limits Event Study Power

Zero statistically significant results from event studies reflect methodological challenges, not necessarily absence of policy effects:

- Only 16 months of real benchmark data vs. 12-month event windows
- Benchmark improvements are discrete (model releases) not gradual
- Future work should use daily/weekly indicators (stock prices, arXiv) or shift to model release events

3. Forecast Convergence Reflects Recent Acceleration

ExogenousAI's 2027 median aligns with AI 2027 Report but deviates from EpochAI (2036) and Bio Anchors (2055):

- Reflects Nov 2025 data showing faster progress than 2023-2024 extrapolations
- Small models (Phi-4-mini 5.6B at 52.8%) achieving SOTA suggests algorithmic efficiency gains
- Exponential extrapolation assumes no slowdown—may be optimistic if scaling laws break

4. Policy Variance Paradox is Mathematically Correct

0% policy contribution arises because all scenarios predict 2027 median:

$$\text{Var}_{\text{policy}} = \text{Var}([2027, 2027, 2027, 2027]) = 0 \quad (10)$$

This is not a bug but reveals that *under current strong trends*, policy affects distribution shapes not locations. More extreme scenarios needed to differentiate central estimates.

4.2 Comparison to Existing Frameworks

EpochAI Direct Approach [Sevilla et al., 2023]:

Strengths: Rigorous scaling law foundations, transparent compute projections, conservative assumptions.

Our Complement: Adds policy scenario branches and external indicator integration. Where EpochAI extrapolates deterministic trends, we model regime changes from policy shocks.

Divergence: ExogenousAI more optimistic (2027 vs 2036) due to recent data showing sustained acceleration.

Biological Anchors [Cotra, 2022]:

Strengths: Provides fundamental bounds via brain-computation comparisons, acknowledges wide uncertainty.

Our Complement: Operationalizes Cotra’s acknowledgment that “transformative AI timelines depend heavily on factors outside the model” by quantifying external policy and economic factors.

Divergence: ExogenousAI significantly more optimistic (2027 vs 2055) due to empirical benchmark data showing faster-than-expected progress. May indicate Bio Anchors’ compute requirements were conservative.

AI 2027 Report [Kokotajlo, 2025]:

Strengths: Rich qualitative scenario analysis capturing race dynamics and coordination failures.

Our Complement: Quantifies scenarios with probabilistic timeline adjustments and confidence intervals, enabling risk assessment.

Agreement: Perfect median alignment (2027) validates observed acceleration trends in 2024-2025.

4.3 Limitations

4.3.1 Data Limitations

1. **Sparse Benchmarks:** Only 16 months of real MMLU-Pro scores limits temporal resolution and event study power
2. **Single Benchmark Focus:** Doesn’t capture multimodal (vision, robotics), scientific reasoning, or long-horizon planning capabilities
3. **Leaderboard Bias:** Excludes proprietary/classified models (e.g., potential unreported Chinese systems)
4. **Retrospective Dating:** Model release dates approximate, not exact evaluation timestamps

4.3.2 Methodological Limitations

1. **Event Study Power:** Insufficient data for robust statistical inference—alternative methods (diff-in-diff, synthetic controls) recommended
2. **Policy Scenario Calibration:** $\pm 5\text{-}10\%$ growth adjustments are illustrative, not empirically validated
3. **Exponential Extrapolation Risk:** Assumes scaling laws continue—vulnerable to regime changes (data exhaustion, compute limits, safety pauses)
4. **Threshold Arbitrariness:** 95% MMLU-Pro as “AGI” proxy, not ground truth of transformative capability

4.3.3 Scope Limitations

1. **Capabilities $\neq$ Deployment:** Timelines measure technical capability achievement, not societal impact or economic transformation
2. **Safety Research Excluded:** Doesn’t model alignment, interpretability, or control problem progress
3. **Economic Factors Simplified:** Compute costs, energy constraints, VC funding treated as static

4. **Black Swans Unmodeled:** Breakthrough discoveries (e.g., new architecture paradigms) or catastrophic failures not captured

4.3.4 Reproducibility Considerations

1. **API Dependencies:** SERP, Alpha Vantage APIs may change pricing/access
2. **Manual Data Entry:** MMLU-Pro leaderboard manually copied (no automated scraper yet)
3. **Proprietary Baselines:** Some literature forecasts use non-public models/assumptions

4.4 Future Work

4.4.1 Short-Term Improvements

1. **Expand Benchmarks:** Add HumanEval+, MATH, GPQA, MMMU for multimodal/reasoning coverage
2. **Automate Data Collection:** Implement HTML scraper for TIGER-Lab Space, daily arXiv API pulls
3. **Denser Indicators:** Use daily stock prices, weekly arXiv counts for event study statistical power
4. **Robustness Checks:** Jackknife resampling, bootstrap confidence intervals for Monte Carlo stability

4.4.2 Medium-Term Extensions

1. **Causal Inference:** Difference-in-differences comparing affected vs. unaffected regions (US vs. EU vs. China benchmark progression)
2. **Synthetic Control Methods:** Construct counterfactual "China without export controls" from weighted combination of unaffected countries
3. **Regime-Switching Models:** Detect structural breaks in trend (e.g., scaling law failures)
4. **Multi-Benchmark Composite:** PCA or factor analysis across MMLU-Pro, HumanEval, MATH for unified capability index
5. **Dynamic Policy Scenarios:** Time-varying adjustments (e.g., gradual compute governance ramp-up)

4.4.3 Long-Term Research Directions

1. **Endogenous Policy Modeling:** Capture feedback loops (progress $\rightarrow$ policy response $\rightarrow$ altered progress)
2. **Alignment Timeline Forecasting:** Parallel predictions for AI safety benchmarks (truthfulness, harmlessness, interpretability)
3. **Economic Impact Models:** Link capability timelines to GDP effects, labor displacement, productivity gains
4. **Multi-Actor Game Theory:** Model US-China AI race dynamics, international coordination incentives
5. **Real-Time Dashboard:** Live integration with Papers with Code API, continuous benchmark ingestion

4.5 Policy Implications

For governance practitioners, this framework offers three actionable insights:

1. Near-Term Preparedness is Critical

Median 2027 AGI timeline with 86.9% probability within 5 years (status quo scenario) suggests:

- Governance institutions need *immediate* capacity building (regulatory frameworks, evaluation protocols, international coordination mechanisms)
- Timelines too short for "wait and see" approaches—must act on current trends
- Safety research community requires urgent scaling to match capability progress

2. Moderate Interventions Affect Probabilities, Not Timelines

Policy scenarios show:

- ± 5 -10% growth adjustments don't shift median years but change probability distributions by 10 percentage points
- Incremental governance (modest compute restrictions, voluntary commitments) affects *tail risks* more than *central estimates*
- More aggressive interventions ($>50\%$ slowdowns) required to meaningfully delay AGI if strong trends continue

3. Uncertainty Monitoring Systems Needed

Variance decomposition reveals:

- 76.9% technical uncertainty dominates—early warning indicators should track benchmark saturation, scaling law deviations
- 23.1% economic uncertainty—monitor semiconductor production, GPU prices, data center capacity
- Real-time dashboards integrating external indicators enable proactive policy adjustment

4.6 Conclusion

This work establishes the feasibility and value of integrating external macroeconomic, geopolitical, and supply chain indicators into AI capability forecasting. Our proof-of-concept implementation demonstrates that:

1. **Strong baseline trends** (+25.5% annual MMLU-Pro growth) dominate moderate policy interventions in near-term forecasts
2. **Data sparsity** limits event study statistical power, indicating need for denser indicators or alternative causal inference methods
3. **Policy scenarios** affect probability distributions (80.8%-90.8%) but not central estimates under current trajectories
4. **Forecast convergence** to 2027 median aligns with recent acceleration, deviating from older literature (EpochAI 2036, Bio Anchors 2055)

The ExogenousAI framework provides foundation for real-time AI progress monitoring systems accounting for exogenous shocks. The complete open-source pipeline (<https://github.com/XAheli/ExogenousAI>) enables replication, extension, and continuous updating as new benchmark data emerges.

The path to transformative AI is not predetermined by scaling laws alone—it will be shaped by policy choices, geopolitical dynamics, and economic constraints. However, our analysis suggests current trends are remarkably resilient to moderate interventions. Policymakers must recognize that under observed +25.5% annual progress rates, preparation timelines are measured in years, not decades. Accurate forecasting must model external factors explicitly while acknowledging that strong technical momentum may overwhelm incremental governance measures.

Appendix A: Implementation Details

Repository Structure

Full code available at: <https://github.com/XAheli/ExogenousAI>

```
ExogenousAI/
  config.yaml          # Configuration (API keys, scenarios)
  requirements.txt      # Python dependencies
  src/
    data_collection/    # MMLU-Pro, policy events, arXiv, stocks
    preprocessing/      # Clean, aggregate, merge (no interpolation)
    analysis/           # Event study, Monte Carlo, meta-forecasting
    visualization/      # Plots for events, scenarios, uncertainty
    run_pipeline.py     # End-to-end execution
  data/
    raw/               # 4,370 policy events, 68 benchmarks, 45mo arXiv
    processed/          # Cleaned datasets
    output/            # Results CSVs, visualizations
  notebooks/           # Jupyter exploratory analysis
```

Data Sources URLs

- **MMLU-Pro Leaderboard:** <https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro>
- **EpochAI Data:** <https://epoch.ai/data>
- **SERP API:** <https://serpapi.com/> (policy events scraping)
- **Alpha Vantage:** <https://www.alphavantage.co/> (NVIDIA stock)
- **arXiv API:** <http://export.arxiv.org/api/query>

Reproducibility Instructions

1. Clone repository: `git clone https://github.com/XAheli/ExogenousAI.git`
2. Install dependencies: `pip install -r requirements.txt`
3. Configure API keys in `.env` file
4. Run pipeline: `python src/run_pipeline.py`

5. Results saved to `data/output/`

Runtime: ~5-10 minutes on standard CPU

Appendix B: Security Considerations

Potential Misuse Risks

1. Adversarial Forecasting

Malicious actors could use this framework to identify optimal timing for disruptive actions (e.g., supply chain sabotage, cyber attacks on semiconductor facilities) that maximize AI development delays.

Mitigation: Restrict detailed supply chain vulnerability mappings to trusted governance channels. Public version aggregates indicators to regional/sector levels.

2. Racing Dynamics Amplification

Public forecasts showing policy-induced delays might incentivize actors to accelerate development before restrictions take effect, creating opposite of intended effect (e.g., "AI arms race" scenario).

Mitigation: Frame scenarios probabilistically rather than deterministically. Emphasize uncertainty and confidence intervals over point estimates. Avoid prescriptive language suggesting specific policy windows.

3. Geopolitical Intelligence Exploitation

Detailed correlations between specific nations' policy actions and capability impacts could be weaponized for competitive advantage in technology competitions.

Mitigation: Aggregate geopolitical indicators to regional levels (East Asia, Europe, North America). Avoid attribution to specific entities or labs.

4. Forecast-Induced Market Manipulation

Timeline predictions could influence semiconductor stock prices, VC funding allocation, or talent migration if widely adopted by financial actors.

Mitigation: Clearly label forecasts as research prototypes, not investment advice. Include prominent disclaimers about limitations and uncertainty.

Limitations Requiring Caution

1. Forecast Accuracy Uncertainty

Event studies establish historical correlations but cannot guarantee future policy impacts follow identical patterns. Black swan events (novel AI paradigms, unexpected breakthroughs) may invalidate extrapolations.

2. Self-Fulfilling Prophecies

Widely adopted timeline forecasts could alter researcher behavior, investment patterns, and policy decisions in ways that validate or invalidate the forecasts themselves (reflexivity not modeled).

3. Indicator Manipulation

If external indicators become known inputs to influential forecasting models, actors might manipulate them (e.g., artificially inflating arXiv submission counts, concealing compute investments) to distort forecasts.

Recommended Safeguards

1. **Ensemble Forecasting:** Always present multiple scenarios rather than single timelines; avoid false precision

2. **Red Team Analysis:** Periodically audit framework for exploitable assumptions or data dependencies
3. **Stakeholder Engagement:** Consult diverse international perspectives to avoid Western-centric policy assumptions
4. **Uncertainty Communication:** Emphasize confidence intervals over point estimates in public communications
5. **Monitoring Feedback Loops:** Track whether forecast publication alters developer/policymaker behavior
6. **Responsible Disclosure:** Coordinate with AI safety organizations before releasing sensitive supply chain analyses

References

- Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.
- Caldara, D. and Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review*, 112(4):1194–1225.
- Chen, M., Tworek, J., Jun, H., et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.
- Cotra, A. (2022). Forecasting Transformative AI: The "Biological Anchors" Method. *Open Philanthropy*. <https://www.openphilanthropy.org/research/report-on-biological-anchors/>
- Dunham, T. J., et al. (2024). Reactor Mk.1 performances: MMLU, HumanEval and BBH. *arXiv preprint arXiv:2406.10515*.
- EpochAI (2023). The Direct Approach. <https://epoch.ai/blog/the-direct-approach>
- EpochAI (2025). Epoch Capabilities Index. <https://epoch.ai/blog/epoch-capabilities-index>
- Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438.
- Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring Massive Multitask Language Understanding. *Proceedings of ICLR 2021*.
- Hendrycks, D., et al. (2021). Measuring Mathematical Problem Solving. *arXiv preprint*.
- Kokotajlo, D. (2025). The AI 2027 Scenario. *AI Futures Project*. <https://ai-2027.com/>
- Lin, X., Deng, G., and Zhang, L. (2025). Integrating Macroeconomic Indicators and U.S. Commodity Futures Prices for Pair Trading on Chinese Commodity Futures. *Proceedings of AAAI 2025 Workshop on AI for Time Series*.
- MacKinlay, A. C. (1997). Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1):13–39.
- Sevilla, J., Heim, L., Ho, A., et al. (2023). Forecasting Transformative AI with the Direct Approach. *Epoch AI Research Report*. <https://epoch.ai/blog/the-direct-approach>

U.S. Department of Commerce (2022). Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items. *Federal Register*, 87 FR 62186.

U.S. Department of Commerce (2024). Export Controls on Semiconductor Manufacturing Equipment. *Federal Register*, 88 FR 73482.