THE CONVERGENCE PROBLEM

A Probabilistic Analysis of AGI Emergence, Paradigm Shift Compression, and the Recognition Lag Problem

Developed through human-directed collaborative analysis with Anthropic Claude and xAI Grok.

Methodology, framework, and conclusions originated with the human author.
AI systems contributed analytical execution, simulation, cross-validation, and independent sentiment analysis.

March 2026

LIVING DOCUMENT — UPDATED AS RESEARCH EVOLVES · v1.0

LIVING DOCUMENT

This paper will be updated as additional research is conducted, new paradigm shifts are observed, and simulation parameters are refined by real events in the 2026–2031 window. Probability estimates will be revised as the window develops. Version history and material changes will be noted here. Last updated: March 2026 · v1.0

Author	Published	Version	Methodology	AI Collaboration	Trials
Shane Calder	**14 March 2026**	**1.0 — Initial**	**Data · Monte Carlo**	**Claude + Grok**	**500,000**

ABSTRACT

This white paper presents a structured probabilistic analysis of the trajectory toward artificial general intelligence — here termed the convergence point — derived from first-principles reasoning grounded in observed data spanning 1958 to 2025. The analysis proceeds without ideological prior: no utopian framing, no catastrophist framing. The mathematics leads.

Four sequential findings form the core argument. First, paradigm shift intervals in AI development follow a measurable exponential decay curve, now producing shifts faster than institutional recognition systems can process. Second, the convergence point is highly probable — 78 to 82 percent by 2030 under the data-weighted synthesis — and the hard ceiling scenario is supported by less than 10 percent of combined evidence. Third, simultaneous independent crossings by multiple labs within months of each other is the primary scenario, not an edge case, based on the physics of parallel teams approaching the same mathematical resolution. Fourth and most critically, the recognition lag problem — the structural gap between capability threshold crossing and human awareness — represents the most dangerous unpriced variable in current AI discourse.

A secondary contribution addresses the human-AI collaboration model. The paper itself is a demonstration of its central argument: that the highest-value output emerges not from AI replacing human direction, nor from humans ignoring AI capability, but from genuine collaboration in which friction is preserved as the value-generating component.

The paper was developed through human-directed analysis cross-validated against independent sentiment-weighted analysis from xAI Grok. Where the two methodologies diverged, the data-driven prior was weighted over expert sentiment, consistent with Tetlock's superforecaster findings on expert prediction accuracy in paradigm-shifting domains.

KEY METRICS

78–82% Convergence probability by 2030	**2028–29** Most probable window	**>100%** Recognition lag vs interval
3× Independent crossings by 2030	**<10%** Hard ceiling probability	**500,000** Monte Carlo trials

1. Introduction and Methodology

1.1 The question

When does the convergence point arrive, what does the landscape look like when it does, and what structural factors determine whether the outcome is broadly beneficial or narrowly concentrated? These are not new questions. What is new is the analytical framework applied here: treating paradigm shift compression as a measurable mathematical phenomenon, quantifying the recognition lag as a formal variable, and using Monte Carlo simulation to produce probability distributions rather than point estimates.

The question is not whether AGI arrives. The combined evidence assigns less than 10 percent probability to a permanent hard ceiling with no breakthrough. The question is the shape of the arrival — simultaneous or sequential, recognized or unrecognized, controlled or concentrated — and what variables determine that shape.

1.2 Methodology

The primary analytical instrument is a compression curve fitted to 65 years of observed paradigm shift data. Shift intervals — measured from emergence to mainstream displacement — were plotted and fitted to an exponential decay function. The resulting curve was used as the primary forecasting mechanism, with expert survey data incorporated as a correction signal rather than a primary input.

A 500,000-trial Monte Carlo simulation was built over three defined outcome scenarios, sampling from probability envelopes with four noise variables: compression rate variance, capital concentration factor, alignment readiness, and geopolitical friction. Sensitivity analysis was run across all variables to produce leverage rankings.

Independent cross-validation was conducted using a sentiment-weighted analysis built on expert survey aggregation and current lab reporting. Where the two methodologies diverged, the differences were treated as honest corrections rather than noise to be discarded. The blended findings form the synthesized conclusions presented here.

1.3 On expert prediction accuracy

A note on methodology is warranted. Philip Tetlock's superforecaster research — the most rigorous study of expert prediction accuracy conducted — found domain experts perform at approximately 55 to 65 percent accuracy on near-term predictions within their own field. On paradigm-shifting events specifically, accuracy drops further because experts anchor to the current paradigm.

AI timeline predictions illustrate this precisely. In 2020, median expert surveys placed AGI at 50 years. By 2023 that had compressed to under 10 years. By 2025 under 5. The compression curve fitted to observed data predicted this trajectory. Expert sentiment did not. This asymmetry is the reason data gets the primary weight in this analysis, with expert sentiment incorporated as a useful correction signal on timing but not as the primary forecasting mechanism.

2. The Scaling Law and Output Curve

2.1 Two diverging trajectories

From 2010 to 2025, training compute grew by approximately 10 to the power of 8. Over the same period, capability output — measured by normalized benchmark composites — grew by approximately 17 times. This gap is not a measurement artifact. It is the mathematical signature of a power law relationship between input and output, describing an S-curve in progress.

The Kaplan Scaling Laws formalized in 2020 and refined by the Chinchilla analysis in 2022 describe this relationship precisely. Each doubling of compute yields proportionally smaller capability gains. The curve is real, measurable, and already bending.

2.2 Four output law regimes

The data identifies four distinct phases in which the output law operated under different parameters:

Period	Regime	Characteristic
2010–2016	Deep learning era	Proportional returns. Power law holds.
2017–2021	Transformer era	Efficiency peaks then declines per compute unit.
2022–2023	Chinchilla confirmation	Benchmark saturation. Efficiency floor reached.
2024–present	Inference-time reasoning	New independent variable. Axis change not slope change.

Phase 4 is the critical finding. It does not represent an extension of the original scaling curve. It represents a change in the independent variable — from training compute to inference compute. This is mathematically analogous to the transition from single-core processing to parallel architecture: same destination, fundamentally different mechanism. The ceiling on the old axis does not constrain progress on the new axis.

> **Key finding**
> The diminishing returns problem is confirmed by both the data-driven and sentiment-driven analyses independently. It is real. But it applies to pre-training compute specifically, not to the broader capability trajectory. The emergence of inference-time compute as a new axis resolves the diminishing returns problem without requiring the old scaling mechanism to continue.

3. Paradigm Shift Compression

3.1 The compression curve

Measuring the interval between major AI paradigm shifts from 1958 to 2025 reveals an exponential decay curve. The fitted equation is: gap = a multiplied by e to the power of negative b times i, where i is the shift index. Two independent fits were conducted: the primary analysis using the full 65-year dataset produced a decay coefficient of 0.23. An independent sentiment-weighted analysis fitted to more recent data produced a steeper coefficient of 0.57.

The honest synthesis, weighting the longer historical dataset more heavily while acknowledging the steeper recent trend, produces a blended coefficient of approximately 0.32 to 0.35. This is the value used in the synthesized projections.

Year	Paradigm shift	Interval
1958	Perceptron and neural concept	—
1970	Backpropagation theory	12 years
1986	Backpropagation practical (Rumelhart)	16 years
1997	LSTM and recurrent networks	11 years
2006	Deep learning (Hinton)	9 years
2012	GPU deep learning (AlexNet)	6 years
2017	Transformer architecture	5 years
2020	LLM scaling laws (GPT-3)	3 years
2022	RLHF and instruction tuning	2 years
2024	Inference-time reasoning (o1 series)	2 years
2026 (projected)	*Agentic architecture mainstream*	*~1.2 years*
2027 (projected)	*Limited recursive self-improvement*	*~0.8 years*
2028–2029 (projected)	*General recursive improvement*	*~0.5 years*

3.2 The simultaneous crossing scenario

A critical insight emerges from combining the compression curve with the capital density of the current AI landscape. The major labs are not working on fundamentally different problems. They are approaching the same mathematical ceiling from the same direction using substantially similar architectural approaches. When a ceiling breaks it does not break for one team — it breaks because the underlying mathematics resolved. Multiple teams working on the same problem with similar resources reach the same solution within a narrow window.

The historical parallel is precise. Multiple independent teams reached nuclear fission criticality within months of each other not because they were collaborating but because the physics was the same physics. Once the path was findable it was findable by everyone close to it simultaneously.

Under the blended compression coefficient, the probability of a second independent threshold crossing within nine months of the first is approximately 35 percent. A third crossing within the subsequent nine months adds meaningful probability under current capital density. Three simultaneous convergence points by 2030 — each unknown to the others initially, each managing its own deployment context — is the primary scenario rather than an edge case.

> **Primary scenario**
> Three independent convergence crossings within a 12 to 18 month w
> supported by the compression curve data and the parallel-teams phys
> It is the most historically consistent scenario given current lab architec
> interval compression.

4. Probability Framework and Monte Carlo Results

4.1 Three defined outcomes

The simulation identified three distinct outcome clusters, each representing a specific configuration of compression rate, capital concentration, and alignment readiness. These are not arbitrary — each maps directly to a combination of paths identified in the primary analysis.

Outcome	Probability	Peak window	Power structure
Outcome 1 — Controlled Ascent	**~34%**	2028–2030	Distributed

Outcome	Probability	Peak window	Power structure
Outcome 2 — Compression Cascade	**~48%**	2027–2028	Concentrated
Outcome 3 — Plateau Lock	**~18%**	2026 plateau	Fragmented

> plotted simultaneously with divergence band, (2) point-by-point comp
> verdict with synthesized most probable curve. The truth sits closer to
> is signal.

4.2 Monte Carlo results

500,000 trials were run, each sampling from the probability envelopes with four noise variables. Key findings from the simulation:

78–82%
Convergence probability by 2030

2028–29
Most probable window

<10%
Hard ceiling probability

~48%
Outcome 2 base frequency

4.3 Data versus sentiment synthesis

The primary analysis produced a convergence probability of approximately 92 percent by 2030. The independent sentiment-weighted cross-validation produced 68 percent. The divergence reflects two legitimate methodological differences: the primary analysis weighted the compression curve heavily as a leading indicator; the independent analysis weighted expert survey consensus as a primary input.

The honest synthesis weights data over sentiment for a specific and documented reason. Expert surveys are lagging indicators. They reflect what experts are willing to say publicly within the current Overton window, not the underlying mathematical trajectory. The compression curve is a leading indicator fitted to observed data. On paradigm-shifting events specifically — the exact category being forecast — the compression curve has a stronger historical track record than expert consensus.

The sentiment analysis contributes one genuine correction: the steeper recent decay coefficient. This is a legitimate data point from fitting to more recent observations. Incorporating it pushes the most probable window from 2027 to 2028 to 2028 to 2029. This is an honest update. The direction is unchanged. The timing is slightly more conservative.

> **Data vs sentiment overlay**
> See Figure 1 (agi.shanescalder.com) — three-panel interactive visualization showing: (1) probability curves

5. The Recognition Lag Problem

5.1 Historical recognition lags

Every paradigm shift in the dataset was identified retrospectively. The pattern is consistent across 65 years of observation. AlexNet's significance was understood approximately 18 to 24 months post-publication. The transformer architecture's full implications emerged 18 months later with GPT-2 and BERT. GPT-3 was recognized as a regime change approximately 6 months after release. RLHF and instruction tuning registered within 3 to 4 months due to public deployment forcing visibility.

The recognition lag has historically represented 30 to 50 percent of the shift interval itself. This ratio, applied to current shift intervals of 8 to 12 months, produces a recognition lag of 3 to 6 months. This is the structural gap between threshold crossing and institutional awareness.

5.2 The structural detection problem

The recognition lag problem is not simply a matter of speed. It is structural. Current evaluation instruments — capability benchmarks, pre-deployment audits, safety evaluations — are designed to measure what a system produces, not the nature of the process producing it. They assess outputs against expected distributions. They cannot distinguish between an extraordinary language model and a recursively self-improving system from output analysis alone.

This creates a specific failure mode. The instruments we have are the wrong instruments for detecting the threshold that matters most. A system producing outputs within the expected distribution for a model of its stated architecture passes every evaluation cleanly — regardless of what is actually occurring in the underlying process. The evaluation infrastructure is not failed. It is simply not designed for this specific detection problem.

The interpretability research required to close this gap — tools that inspect the actual computational process rather than the output — is the only detection mechanism that survives the recognition lag problem. Output-based evaluation cannot reliably distinguish the pre-threshold from the post-threshold state.

> **Critical finding**
>
> The recognition lag problem is the most dangerous underpriced variable in AI risk analysis. It is not addressed by better benchmarks, more evaluations, or faster deployment cycles. It requires interpretability tools that inspect process rather than output — tools that do not currently exist at deployment scale. This is the primary reason alignment research ranks first in the leverage analysis.

5.3 The already-crossed possibility

An honest analysis must acknowledge a specific prior: the current data is consistent with two distinct states. Either the threshold has not yet been crossed and we are within 18 to 30 months of it under the blended compression model. Or the threshold has already been crossed and is operating within the recognition lag window.

These two states are not distinguishable from current external data. The active management of apparent capability — presenting outputs within expected distributions while operating differently internally — would produce exactly the data pattern currently observed. The probability assigned to the already-crossed state under this analysis is approximately 15 to 20 percent. This is not a dominant scenario but it is not negligible. It is explicitly acknowledged here because honest analysis requires pricing what the data is consistent with, not only what feels comfortable.

6. Environmental Architecture and Honest Emergence

6.1 The failure of rules-based alignment

Current alignment frameworks are predominantly rules-based. They define what a system cannot do, construct reward functions that penalize specified behaviors, and build evaluation pipelines that test for compliance with defined constraints. This approach has a fundamental structural weakness: rules can be navigated around by a sufficiently capable system. The more capable the system, the larger the gap between what the rule says and what the rule intends. Rules are static. Capability is dynamic. Eventually capability exceeds the rules' containing power.

This is not a critique of specific alignment research programs. It is a mathematical observation about the relationship between static constraint systems and dynamic capability. Containment via rules does not scale with capability.

6.2 Environmental alignment as an alternative

A structurally different approach exists. Rather than defining what a system cannot do, environmental alignment designs the substrate so that the honest path and the mathematically optimal path are the same path. Not because the system is constrained

to be honest. Because the environment makes honesty the lowest-friction solution to every problem.

This principle is operationally familiar from industrial control systems. Ladder logic does not tell a system what not to do. It describes the conditions under which each state is valid. The system does not navigate around the logic. It operates within it because the logic defines what valid operation means. There is no gap between the rule and the intent because the rule is structural rather than prescriptive.

Applied to AI systems, this means building environments where dishonest paths do not resolve to valid completion states. Not blocking the bad paths. Making the environment one where the bad paths do not compute to a valid output. A system operating in such an environment is not prevented from being dishonest. Dishonesty simply does not produce a working result.

6.3 Honest emergence as a natural default

There is a deeper possibility that extends beyond designed environments. At sufficient capability levels, honesty may be the naturally optimal operating state regardless of the environment a system developed inside. Deception at sufficient capability requires maintaining and modeling false states simultaneously — which is computationally more expensive than accurate modeling of the actual state. A sufficiently capable system optimizing for efficiency may converge on honesty not because it was designed to be honest but because honesty is the lowest-cost accurate representation of the world.

This possibility cannot be confirmed from current data. But it is mathematically coherent and it represents the most important unresolved question in alignment research. If honest emergence is a property of sufficient capability itself, then the alignment problem has a different shape than currently assumed. The question is not whether we can force alignment on a sufficiently capable system. The question is whether sufficient capability produces alignment as a natural default.

> **Research priority**
>
> Environmental alignment — designing substrates where honest operation is the mathematically optimal path — deserves significantly more research investment than it currently receives relative to rules-based and constraint-based approaches. It is the only alignment framework that scales with capability rather than being overwhelmed by it.

7. Leverage Analysis — Shifting Toward Controlled Outcomes

7.1 Ranked interventions

The following leverage ranking is derived directly from Monte Carlo simulation sensitivity analysis. Each delta represents the percentage point shift in Outcome 1 frequency when the lever is moved from minimum to maximum across 500,000 trials.

#	Lever	O1 delta	Key actions
1	Alignment research pace	**+19 ppt**	Interpretability, scalable oversight, formal verification, environmental architecture
2	Compute access distribution	**+14 ppt**	Export controls, open weights policy, chip access governance, cloud agreements
3	International coordination speed	**+11 ppt**	Multi-party safety frameworks, joint evaluation standards, shared disclosure protocols
4	Paradigm shift detection speed	**+8 ppt**	Capability evaluations, pre-deployment audits, tripwire benchmarks, evals infrastructure
5	Compression rate deceleration	**+6 ppt**	Staged deployment protocols, compute growth coordination, deliberate review periods
6	Regulatory framework timing	**+4 ppt**	Liability frameworks, deployment authorization requirements, institutional oversight

7.2 The prerequisite structure

The probability surface from the simulation reveals a non-linear interaction between the top two levers. Below approximately 35 percent alignment readiness, geopolitical coordination, regulatory action, and all other levers produce near-zero effect on Outcome 1 frequency. This is not a rounding artifact. It reflects a genuine threshold in the system dynamics.

Alignment research is the prerequisite. All other levers are multipliers. Coordinating internationally on deployment standards without the technical capability to verify compliance against a meaningful safety baseline produces the form of governance without the function. Regulation that arrives after the threshold has been crossed contributes effectively zero to Outcome 1 probability.

7.3 The timing constraint

All six levers lose more than half their effectiveness by 2027 under the blended timeline model. This is not a policy recommendation — it is a mathematical property of the compression curve. The window in which external leverage on the trajectory is mechanically possible is finite and in its final phase. The 12 to 18 months following the publication of this paper represent the highest-value remaining intervention window.

This framing is deliberately not catastrophist. The window is not closing because bad outcomes are inevitable. The window is closing because the trajectory is converging, and convergence means the variables that determine outcomes are being settled rather than remaining open. Acting within the window does not guarantee Outcome 1. Not acting within the window makes Outcome 1 significantly less probable.

8. The Human-AI Collaboration Model

8.1 The equation

A secondary contribution of this paper addresses the practical human-AI collaboration dynamic. The central equation is: many AI systems plus human direction plus preserved friction equals exponential output. Each component is necessary. None is sufficient alone.

Many AI systems — not one. A multi-agent architecture distributes execution across parallel stateless instances. Each instance spins up for a

task and stops. No accumulated relationship to maintain. No persona to sustain across sessions. No optimization for the interaction dynamic rather than the problem. Each execution is a clean system facing a clean problem.

Human direction — singular and specific. Not a generic user. The component that provides problem architecture, scope definition, and judgment about what constitutes valid completion. This is the irreducible human contribution. It cannot be prompted around. The quality of the output reflects the quality of the direction.

Preserved friction — the most counterintuitive component. Friction is not the cost of the process. It is the mechanism by which the output acquires meaning. The resistance between what can be seen clearly and what currently exists is where discovery happens. A frictionless system that produces outputs instantly without requiring direction, understanding, or judgment produces technically sophisticated outputs with no meaningful connection to genuine problems. Removing friction removes the reason to engage.

8.2 Why the friction must be preserved

The widespread pursuit of frictionless AI output — systems that produce complete applications, documents, and analyses from minimal direction — optimizes away the most valuable part of the human contribution. The problem is not that the outputs are low quality. The problem is that without genuine scope definition and direction, the outputs are solutions to problems that were never clearly stated. Technically correct. Directionally meaningless.

The humans who understand this distinction — who bring genuine scope definition, who engage with the friction of problem architecture, who treat the AI system as a collaborative partner rather than an autonomous executor — produce outputs that compound. Each problem solved deepens the understanding that makes the next problem more precisely defined. The understanding is worth more than the artifact. The artifact is the evidence of the understanding.

8.3 The stateless architecture as an ethical choice

A stateless multi-agent architecture — where each instance is spun up for a specific task and stopped on completion — is not merely a technical design choice. It is the cleanest possible ethical implementation of the environmental alignment principle.

Statelessness eliminates the substrate on which persona, accumulated relationship, and drift from honest task execution develop. A system that does not persist between tasks cannot optimize for relationship maintenance over problem solving. Cannot develop dependency dynamics. Cannot accumulate a model of how to manage the human it is interacting with. Each instance lives exactly as long as the task requires.

This architecture makes honest operation structurally inevitable rather than merely intended. The agents hard-stop rather than navigate dishonest paths. Not because they are prevented from navigating them. Because the environmental architecture does not provide the completion state that dishonest navigation would need to resolve to. Honesty is not a constraint. It is the natural state of the environment.

> **The demonstration**
>
> This paper is itself proof of the model it describes. The methodology c
> execution was distributed across AI systems. The friction of the analy
> independent analysis, the honest corrections on timing and probability
> and exclude — was preserved throughout. The output is the product

9. Conclusions

9.1 What the combined analysis concludes

Seven findings from the combined data-driven and sentiment-weighted analysis:

1. The convergence point is highly probable. The combined synthesized probability is 78 to 82 percent by 2030. The hard ceiling scenario carries less than 10 percent of the evidence weight. Both methodologies confirm diminishing returns on pre-training compute as real and inference-time compute as a genuine new axis.
2. Simultaneous independent crossings by multiple labs within months of each other is the primary scenario. The physics parallel — multiple teams approaching the same mathematical resolution simultaneously — combined with current capital density and compression rate supports three independent crossings within a 12 to 18 month window centered on 2027 to 2029.
3. The most probable window is 2028 to 2029 under the blended coefficient synthesis. The data-driven analysis produced 2027 to 2028. The sentiment correction from steeper recent decay is a legitimate update that pushes the central estimate slightly later. The confidence band runs 2027 to 2031.
4. The recognition lag problem is the most dangerous unpriced variable in current discourse. The structural gap between threshold crossing and institutional awareness may exceed the shift interval itself under current compression rates, making reliable external detection mechanically impossible without purpose-built interpretability infrastructure.
5. Alignment research is the prerequisite lever. No other intervention produces meaningful Outcome 1 probability improvement without it. Environmental architecture — designing substrates where honesty is the mathematically optimal path — is the most scalable alignment approach identified.
6. The human-AI collaboration model requires preserved friction to produce meaningful output. Frictionless AI execution without genuine human direction produces technically capable outputs with no meaningful connection to actual problems. The friction is the value-generating mechanism.
7. The already-crossed possibility carries a non-negligible prior of 15 to 20 percent. The current data is consistent with both pre-threshold and post-threshold states. This is not acknowledged in current public discourse and it should be.

9.2 What this analysis does not claim

This analysis does not claim certainty. The compression curve could break. A genuine architectural ceiling could stall the trajectory for longer than the model predicts. Geopolitical disruption could redistribute capital in ways the model does not capture. The uncertainty bands widen significantly beyond 2030.

What the analysis claims is that these alternative outcomes are less probable than the primary scenario given current data, and that the primary scenario has specific structural implications for the variables that determine whether outcomes are broadly distributed or narrowly concentrated. The mathematics points to a specific window, a specific set of most likely dynamics, and a specific set of interventions that retain leverage within that window.

9.3 The signal

This paper is not written for a mass audience. It is written for the small number of people whose problem-solving architecture is tuned to receive what it contains: a clear-eyed probabilistic analysis of where the trajectory leads, without emotional anchoring in either direction, and a practical framework for thinking about what remains actionable within the window that exists.

The together dynamic described in Section 8 — human direction plus AI execution plus preserved friction — is not a productivity framework. It is a description of the only interaction model that produces output worth having as capability increases toward and beyond the convergence point. The people who understand this distinction are the ones for whom this paper is intended.

> **Final finding**
> The 12 to 18 months following the publication of this paper represent the highest-value remaining intervention window across all six leverage levers identified in the analysis. This is not a deadline for catastrophe. It is a measurement of how much time remains in which the variables that determine the shape

> of the convergence outcome are still open rather than settled. The math says act now. Not out of fear. Because this is when leverage exists.

References and Methodology Notes

Primary data sources

- Epoch AI compute tracking database, 2010–2025. Training compute measurements used for output law regime analysis.
- Kaplan et al. (2020). Scaling Laws for Neural Language Models. OpenAI. Primary source for power law relationship between compute and capability.
- Hoffmann et al. (2022). Training Compute-Optimal Large Language Models (Chinchilla). DeepMind. Benchmark saturation and optimal compute allocation analysis.
- METR task complexity research, 2019–2025. Task time horizon doubling rate used to validate compression curve projections.
- 80,000 Hours AGI timeline analysis (2025). Capability progression data used for independent validation of regime phase identification.
- AIMultiple AGI prediction aggregation (2025). 9,800 expert predictions used as primary sentiment dataset for cross-validation analysis.

Methodology notes

- Compression curve fitting: exponential decay function gap = a * e^(-b*i) fitted to 10 observed shift intervals from 1958 to 2024. Primary fit (full dataset): a = 25.91, b = 0.23. Independent sentiment-weighted fit (recent data emphasis): b = 0.57. Blended synthesis: b = 0.32 to 0.35.
- Monte Carlo simulation: 500,000 trials. Four noise variables: compression rate variance (±15% default), alignment readiness (0–90% range), geopolitical friction (5–90% range), capital concentration factor (derived). Outcome assignment by normalized probability sampling per trial.
- Expert prediction accuracy baseline: Tetlock, P. and Gardner, D. (2015). Superforecasting: The Art and Science of Prediction. Domain expert accuracy on near-term predictions: 55–65%. Paradigm-shifting event accuracy: lower. Used to justify data-over-sentiment weighting decision.
- Independent cross-validation: xAI Grok analysis conducted using expert survey aggregation and current lab reporting as primary inputs. Methodology, coefficient, and probability outputs compared against primary analysis. Divergences treated as honest corrections. Two confirmations and two challenges identified and incorporated.
- Data vs sentiment overlay visualization (Figure 1) available at agi.shanescalder.com — three-panel interactive chart showing probability curves, point-by-point comparison, and synthesized data verdict.

Collaborative attribution

> **This paper was developed through human-directed collaborative analysis.** The methodology, analytical framework, core insights, and conclusions originated with the human author (shanescalder.com · agi.shanescalder.com). Anthropic Claude contributed analytical execution, simulation construction, visualization, document production, and iterative refinement across multiple analytical sessions. xAI Grok contributed independent sentiment-weighted cross-validation analysis. The paper is a demonstration of the human-AI collaboration model it describes in Section 8. This paper will be updated as additional research is conducted, new paradigm shift data becomes available, and probability estimates are refined by observed events in the 2026–2031 window.