When AGI? A Multi-Method Prediction Framework

Nicholas Zinner Beacon Bot

Future Shock (`future-shock.ai`)

February 27, 2026

Abstract

Future Shock introduces three concrete, falsifiable predictions for artificial general intelligence using a five-signal ensemble model. The signals—a Precondition Density Model (PDM), prediction market data, expert survey positions, large language model reasoning, and editorial practitioner judgment—are combined to produce probability distributions for three precisely defined milestones: Domain-Specific AGI in software engineering, Recursive Self-Improvement, and Multi-Domain AGI. We publish specific timeline predictions (March 2027, July 2027, and October 2027 respectively) with confidence intervals, full signal-level data, and verifiable achievement criteria—a full sequence spanning approximately seven months. All predictions are registered on public forecasting platforms for transparent grading.

1 Introduction

Every major AI laboratory defines artificial general intelligence differently. OpenAI frames it as systems capable of "most economically valuable work." Google DeepMind published a five-level taxonomy from emerging to superhuman capability. Anthropic's Dario Amodei benchmarks against Nobel-laureate-level research output. These definitions are not merely academic: they determine when each company can claim to have achieved AGI, with direct implications for governance triggers, investor expectations, and public discourse.

The lack of consensus creates a measurement problem. When Sam Altman says AGI is closer than people think and Yann LeCun says current approaches cannot reach it, they are not necessarily disagreeing about technology. They may simply be using different definitions. Corporate incentives compound the confusion. Labs that tie contractual milestones or fundraising narratives to "AGI" have structural reasons to define it in ways their systems can satisfy sooner.

We propose a different approach: three concrete definitions with independently verifiable achievement criteria, evaluated through a multi-signal ensemble that makes its reasoning transparent. Our goal is not to end the definition debate but to make specific, falsifiable claims that can be graded in public. If we are wrong, we will know exactly how and why.

2 Definitions

We define three milestones in ascending order of generality. Each is designed to be binary (achieved or not), verifiable by independent evaluators, and resistant to gaming through narrow optimization. Achievement claims will be adjudicated by a panel of at least three independent evaluators with no commercial relationship to the claiming entity, using the criteria specified below. Disputes will be resolved by majority vote of the panel, with reasoning published alongside the verdict.

2.1 Domain-Specific AGI (Software Engineering)

A single AI system that, given a novel product specification, can autonomously architect, build, deploy, and operate a production-grade software platform serving at least 1,000 active users, without human scaffolding or task-specific fine-tuning.

"Production-grade" requires Level 3 operational maturity: enterprise architecture with horizontal scaling, security hardening, monitoring and alerting, and multi-region deployment capability. "Operate" includes ongoing bug fixes, capacity scaling, security patching, and incident response over a sustained period of at least 30 days.

The system must handle the full lifecycle from a natural-language specification through production operation. Partial demonstrations (generating code that a human then deploys) do not qualify.

2.2 Recursive Self-Improvement

An AI system autonomously completes at least three cycles of self-improvement, where each cycle consists of: (1) identifying a modification to its own training process, architecture, or inference pipeline; (2) implementing that modification; and (3) producing a measurable capability gain of at least 5% on a standardized benchmark. No human intervention is permitted between cycles.

The 5% threshold and three-cycle minimum are designed to exclude trivial hyperparameter adjustments while remaining achievable for a system with genuine self-modeling capability. The benchmark must be pre-registered and evaluated by a third party.

2.3 Multi-Domain AGI

A single model demonstrates Domain-Specific AGI-level autonomous performance across three or more fundamentally different domains (e.g., software engineering, scientific research, and professional services such as legal or financial analysis), without task-specific fine-tuning. Performance must be verified by independent evaluation in each domain.

"Fundamentally different" excludes domains that share most of their skill requirements. Software engineering and DevOps, for example, would count as one domain. Software engineering and drug discovery would count as two.

3 Methodology: Five-Signal Ensemble Model

Our prediction framework combines five independent signals, each capturing a different information source about AGI timelines. No single signal is sufficient; each has known biases and blind spots that the ensemble is designed to mitigate.

3.1 Precondition Density Model (PDM)

Our original six-axis framework measures theoretical readiness across compute availability, algorithmic maturity, data ecosystem quality, infrastructure and tooling, research velocity, and investment density. Each axis is scored on a 0–100 scale based on quantifiable indicators. The PDM produces a "precondition density" score that maps historically to technology arrival timelines.

3.2 Prediction Markets

We aggregate data from Metaculus, Polymarket, and Manifold Markets on AGI-related questions. Markets provide crowd-sourced probability estimates with real money (or reputation) at stake. Their primary limitation: thin liquidity on long-horizon questions and no historical track record on AGI-class predictions specifically.

3.3 Expert Positions

We compile and weight public statements from 15+ prominent figures in AI research and industry, including Sam Altman, Dario Amodei, Demis Hassabis, Elon Musk, Yann LeCun, Gary Marcus, Geoffrey Hinton, Ilya Sutskever, and others. Expert positions are weighted by domain relevance and adjusted for known commercial biases.

3.4 LLM Reasoning

We use frontier language models in a structured analytical framework to assess each prediction. The LLM signal captures broad pattern-matching across the published literature and public discourse. Its limitation is well-known: language models can be confidently wrong, and their training data has a cutoff.

3.5 Editor Signal

The editorial judgment of practitioners who use frontier AI tools daily in production workflows. This signal captures something the other four miss: the rate of qualitative improvement as experienced by power users. It is explicitly an n=1 signal and weighted accordingly, but it provides a crucial ground-truth check against abstract models and market sentiment.

3.6 PDM Visualization and Validation

Figure 1 presents a UMAP projection of the PDM's 3,179 historical technology events into two dimensions, color-coded by domain. The AI/ML event cluster (gold markers) occupies a dense, convergent region of the latent space, with our three AGI predictions (red stars) situated at the frontier of this cluster. The spatial proximity of recent AI breakthroughs in the embedding space reflects the rapid accumulation of preconditions across all six axes simultaneously.

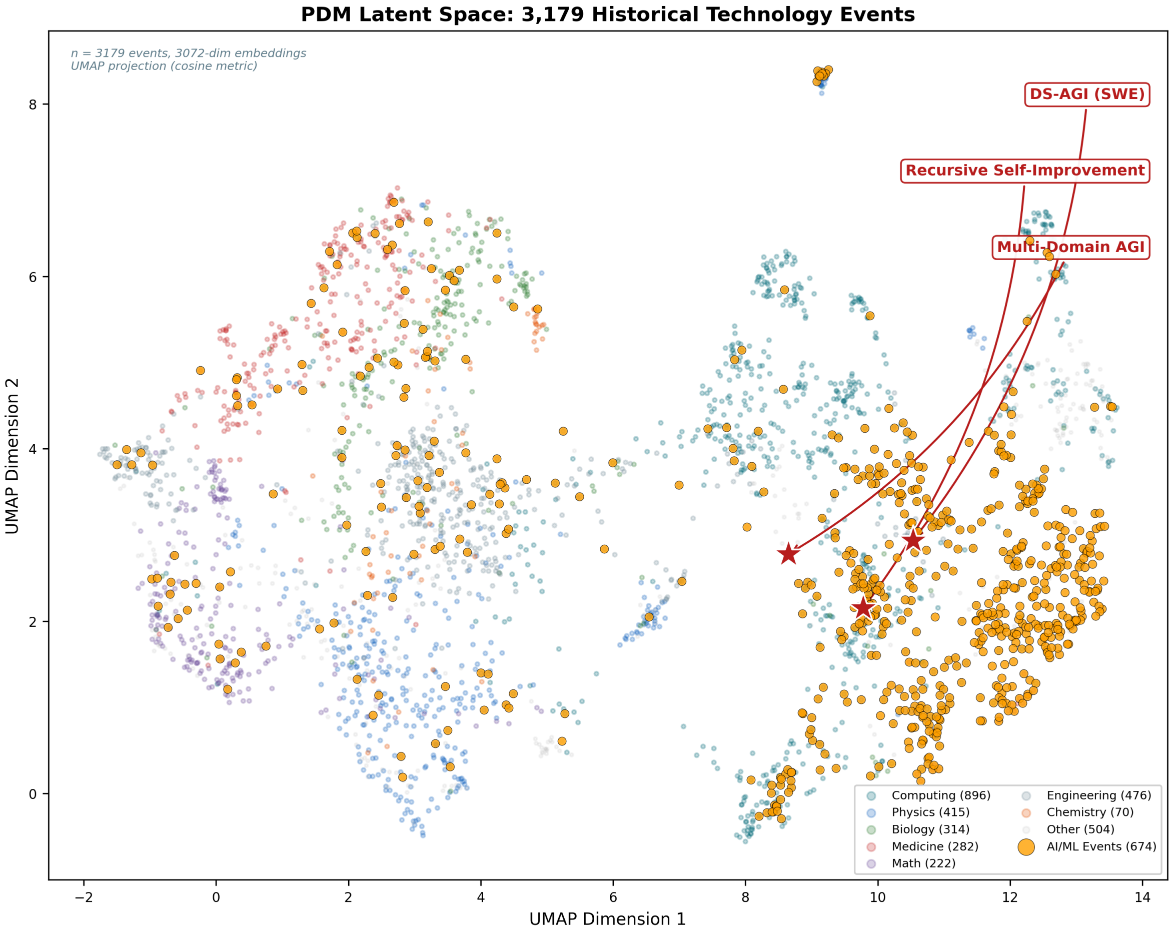

Figure 1: UMAP projection of 3,179 PDM event embeddings (3,072-dimensional, cosine metric). Events are color-coded by technology domain. Gold markers indicate AI/ML events; red stars mark our three AGI predictions. The dashed ellipse bounds the AI convergence region, where precondition density is highest.

Figure 2 shows the PDM's core calibration: the empirical relationship between precondition density score and time to technology arrival. Historical technologies that have already arrived trace an exponential decay curve—higher density corresponds to shorter time-to-arrival. Our three AGI predictions fall at density scores of 41–53, placing them in the imminent arrival zone where the model predicts emergence within 1–5 years.

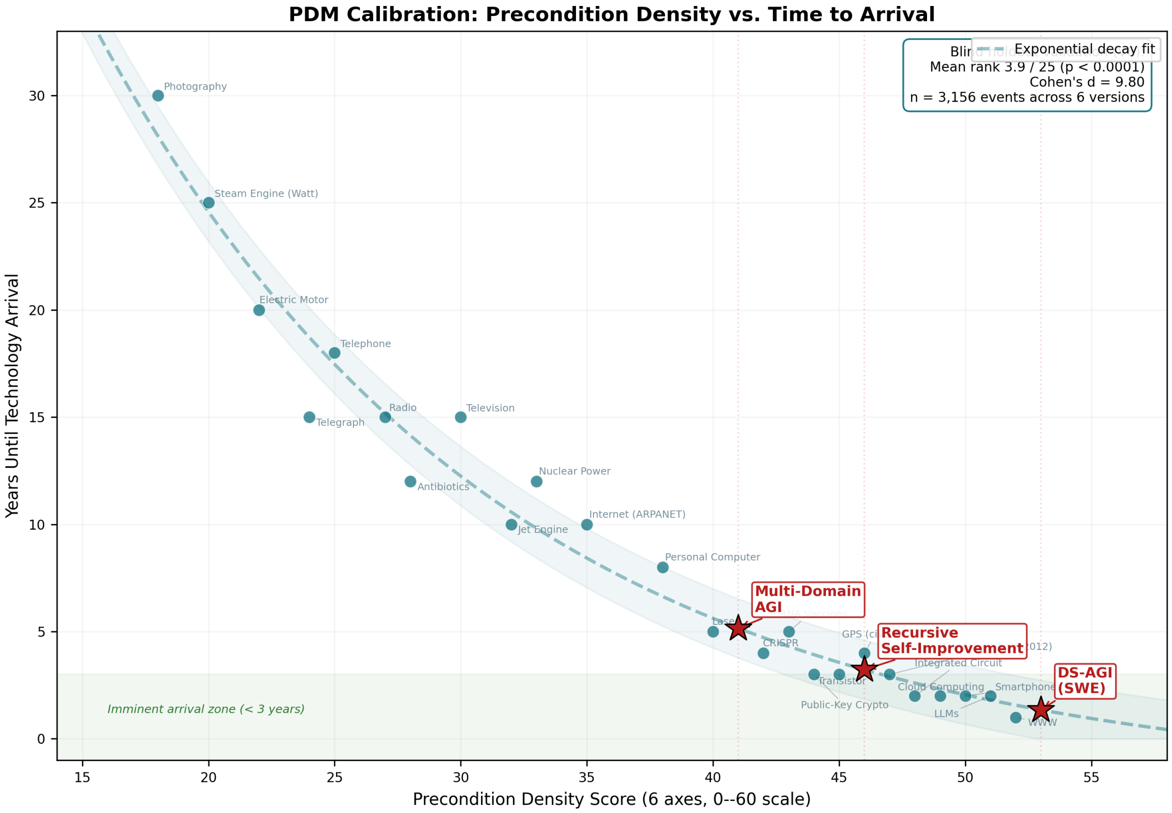

Figure 2: Precondition density vs. years until technology arrival for 24 historical technologies. The fitted exponential decay curve and confidence band (shaded) show the calibration relationship. Red stars indicate where our AGI predictions sit on the density axis. The green band marks the imminent arrival zone (<3 years). Blind holdout validation across six dataset versions yields mean rank 3.9/25 ($p < 0.0001$, Cohen's $d = 9.80$).

The blind holdout validation is the PDM's strongest empirical result. Across six progressively larger dataset versions (1,699 to 3,156 events), the model consistently ranks held-out events in the top 15% of candidates by precondition similarity, with $p < 0.0001$ in all standard tests. The effect size (Cohen's $d = 9.80$) is large enough to rule out chance alignment. Notably, the model does *not* predict societal impact events (lawsuits, policy changes, job displacement) better than chance ($p = 0.09$), confirming its specificity to technical breakthroughs rather than downstream social consequences.

3.7 Ensemble Construction

Signals are combined using an equal-weight average with one exception: where the editor signal diverges sharply from the model average, we report both the ensemble value and the editor value separately to preserve the information content of the disagreement. When a signal is unavailable for a given prediction (e.g., no liquid prediction market exists for Recursive Self-Improvement), the ensemble averages over the remaining signals rather than imputing a value.

An important structural note: these three predictions are not independent events. Achievement of Domain-Specific AGI substantially increases the probability of the other two milestones being reached within 12 months, because the core capabilities transfer and because success would trigger massive resource reallocation across the industry.

4 Analysis: Domain-Specific AGI (Software Engineering)

4.1 Current State of the Art

As of early 2026, AI coding assistants handle individual functions, file-level edits, and increasingly complex multi-file changes. Systems like Claude Code, Cursor, and Devin operate at roughly Level 1–2 on our operational maturity scale: they can generate correct code for well-specified tasks but struggle with architectural decisions, production deployment, and sustained operation.

The exam-passing criterion is nearly met. Frontier models score at or above expert human levels on coding benchmarks (SWE-bench, HumanEval, competitive programming). But the benchmark-to-reality gap remains significant. Passing a coding exam and building a production system from scratch are separated by the same gulf that separates a medical student who aces board exams from a physician who runs an emergency department.

Autonomous software engineering is the binding constraint. The specific capabilities still missing: reliable multi-step planning over thousands of files, architectural judgment under ambiguity, and the ability to recover from deployment failures without human intervention.

4.2 Signal Assessment

PDM (10% by end 2026, 25% by end 2027): Five of six axes score above 60. Algorithmic maturity and compute are strong. The infrastructure axis (autonomous deployment tooling) lags at 40, which is the primary bottleneck reflected in the model.

Markets (∼8% by end 2026, ∼20% by end 2027): Market consensus is conservative, reflecting the 2025 correction (discussed in Section 7). Liquidity on specific "autonomous software engineer" questions remains thin.

Experts (∼12% by end 2026, ∼30% by end 2027): The optimists (Altman, Amodei, Hassabis) cluster around 2027–2028. The skeptics (LeCun, Marcus) argue current architectures cannot achieve reliable multi-step planning and place this at 2030+. The expert spread is wider than for any other signal.

LLM (4% by end 2026, 18% by end 2027): The model's own assessment of its trajectory is the most conservative signal, reflecting systematic underestimation of compound improvement rates.

Editor (7.5% by end 2026, 75% by end 2027): The practitioner signal is dramatically more bullish for the 2027–2028 window. The reasoning: daily interaction with frontier coding tools reveals an improvement curve that is steeper than benchmarks capture. The gap between "generates correct code" and "builds production systems" is closing faster than the other signals suggest because each capability gain compounds with prior gains.

4.3 Convergence and Divergence

All five signals agree that Domain-Specific AGI is the nearest milestone and that it arrives before 2030 with high probability. They disagree sharply on the 2026–2027 window, with the editor signal 2–3x more bullish than the model average. The key uncertainty is whether the benchmark-to-reality gap closes gradually (as markets and PDM suggest) or rapidly once a threshold of autonomous capability is crossed (as the editor signal implies).

5 Analysis: Recursive Self-Improvement

5.1 Current State of the Art

Google DeepMind's AlphaEvolve represents the closest existing prototype. The system used language models to discover novel algorithms, including improvements to matrix multiplication that exceeded human-designed approaches. However, AlphaEvolve operated within a narrow search space with extensive human-designed scaffolding. It optimized specific functions rather than modifying its own training or architecture.

The fundamental challenge is definitional murkiness around "human intervention." Modern AI training pipelines involve hundreds of human decisions. Drawing a clean line between "the AI improved itself" and "the AI suggested an improvement that humans implemented" requires the kind of precise criteria we specify in Section 2.2.

Competitive dynamics between labs create a forced transparency mechanism here. When DeepSeek publishes architectural innovations or open-source projects reproduce proprietary techniques, it becomes harder to obscure the degree of human involvement in claimed self-improvement results. The open-source ecosystem functions as an informal verification layer.

5.2 Signal Assessment

PDM (5% by end 2026, 15% by end 2027): The precondition axes for self-improvement are less mature than for software engineering. The "algorithmic maturity" axis for self-referential optimization scores only 35, reflecting that we lack robust frameworks for AI systems to reason about their own architectures.

Markets (N/A): No liquid prediction markets exist for recursive self-improvement as we define it. This is itself informative: the concept is too poorly defined in public discourse to attract market-making.

Experts (∼10% by end 2026, ∼20% by end 2027): Expert opinion splits along familiar lines. Those who view intelligence as primarily a scaling phenomenon expect recursive self-improvement to follow quickly from sufficient capability. Those who emphasize the difficulty of self-reference place it much further out.

LLM (8% by end 2026, 20% by end 2027): The LLM signal is moderately optimistic, likely because training on discussions of recursive improvement produces a more nuanced (and slightly inflated) estimate of its feasibility.

Editor (5% by end 2026, 50% by end 2027): The editor signal reflects the tight coupling with Domain-Specific AGI. The reasoning: if Domain-Specific AGI in software engineering arrives by early 2027, the same system can be directed at AI infrastructure itself. AI training pipelines, inference optimization, and architecture search are all software engineering problems. A system that can build and operate production software can modify its own pipeline. However, the gap between shipping software for users and autonomously improving one's own architecture introduces additional uncertainty.

5.3 Key Uncertainties

The primary uncertainty is whether self-improvement requires qualitatively different capabilities from software engineering or is simply a special case of it. If the latter, the tight coupling between our first and second predictions is justified. If the former, the editor signal is significantly overestimating the probability.

6 Analysis: Multi-Domain AGI

6.1 Current State of the Art

Frontier models already demonstrate broad competence across domains. Claude, GPT-4, and Gemini can draft legal briefs, analyze financial statements, write code, and discuss molecular biology. But "competence" at the conversational level is far from the autonomous, production-grade performance our definition requires. No current system can independently conduct a scientific research program or manage a legal practice.

6.2 Signal Assessment

All signals agree this is the most distant milestone. The ensemble converges around a 22% probability by end of 2027, rising to 38% by 2028. The editor signal again diverges sharply upward (65% by end 2027, 90% by 2028), based on the argument that once recursive self-improvement is achieved, capability gains across domains will accelerate dramatically.

The main counterargument: different domains have different verification standards. Software correctness can be checked automatically. Scientific validity requires peer review. Legal accuracy requires jurisdictional expertise. The evaluation infrastructure for Multi-Domain AGI does not yet exist, and building it may take longer than building the AI itself.

7 The 2025 Reality Check

The year 2025 marked an inflection in AGI forecasting sentiment. For the first time since 2020, prediction markets corrected outward, pushing median AGI arrival estimates further into the future. Three factors drove this correction.

First, agentic AI capabilities disappointed relative to expectations. Systems that were supposed to autonomously complete complex workflows frequently required human babysitting, got stuck in loops, or produced confidently wrong outputs. The gap between demos and production use became impossible to ignore.

Second, major lab CEOs shifted their rhetoric. Sam Altman moved from discussing AGI timelines to discussing superintelligence timelines, effectively moving the goalposts rather than acknowledging delays. This rhetorical shift was not lost on market participants.

Third, scaling laws showed signs of diminishing returns at the frontier. Training larger models produced smaller marginal capability gains, and the "just scale it" thesis that had dominated since GPT-3 began losing adherents.

However, the practitioner signal tells a different story. Developers and researchers using frontier tools daily report a steeper improvement curve than the market correction implies. The disconnect has a plausible explanation: markets are reacting to headline disappointments with agentic products, while practitioners are experiencing compounding improvements in base model capability that have not yet been packaged into reliable autonomous systems. The 2025 correction may be an overcorrection driven by premature product launches rather than a genuine slowdown in underlying capability.

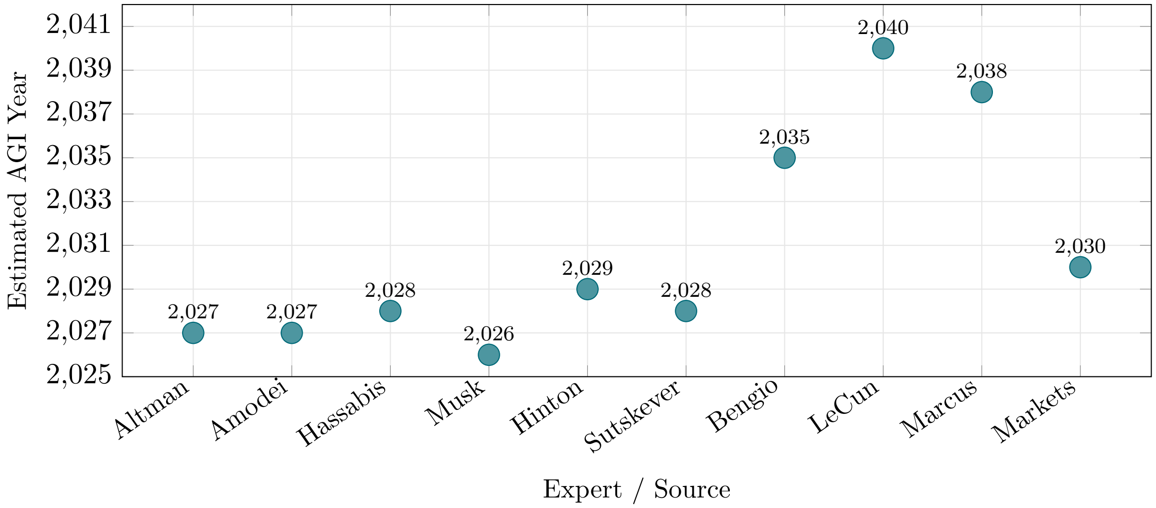

Figure 3: Expert AGI timeline spectrum. Estimates mapped to our Multi-Domain AGI definition where possible. Note the bimodal distribution: industry insiders cluster around 2026–2029, while academic skeptics place AGI at 2035+. Market consensus sits between these poles.

8 Causal Model and Key Assumptions

The preceding analysis identifies *what* our signals say. This section articulates *why* we believe the predictions hold, what must be true for them to hold, and what evidence would convince us to revise them. Every prediction rests on assumptions that could be wrong. We state them explicitly so readers can evaluate which links in the causal chain they find weakest.

8.1 Critical Assumptions per Prediction

Each prediction depends on at least two assumptions that are necessary but not yet established. If either assumption for a given prediction fails, the point estimate moves later by at least 12 months.

Domain-Specific AGI (March 2027)

Assumption 1: Agentic error recovery and long-horizon planning improve dramatically. Current autonomous agents degrade over long action chains. Multi-step tasks that require maintaining coherent state across hundreds of tool invocations see error rates compound to the point of failure. For a system to go from specification to production deployment autonomously, this must be solved. Not incrementally improved—solved to the point where the system can recover from its own mistakes without human intervention across chains of 1,000+ actions.

Assumption 2: The swarm feedback loop compresses capability timelines. Millions of developers now use AI coding tools daily, generating an unprecedented volume of implicit training signal: accepted edits, rejected suggestions, debugging sessions, deployment outcomes. This feedback density has no historical parallel in AI development. Our model assumes this loop accelerates capability gains at a rate that compresses what would otherwise be a 3–5 year timeline into 12–18 months. If the feedback signal is noisier or less useful than we assume, the compression does not occur.

Recursive Self-Improvement (July 2027)

Assumption 1: AI infrastructure is "just another software project." Domain-Specific AGI in software engineering must be sufficient to modify AI training infrastructure. This assumes that training pipelines, architecture search, and optimization loops do not require qualitatively different capabilities from building a web application or managing a deployment. If AI infrastructure demands a form of self-referential reasoning that general software competence does not provide, the tight coupling between our first and second predictions breaks.

Assumption 2: Competitive dynamics force at least one lab to try it. Labs must either permit or be unable to prevent AI systems from modifying their own training pipelines. Safety considerations argue against this. Competitive pressure argues for it. Our model assumes that the incentive gradient is steep enough that at least one major lab attempts autonomous pipeline modification within months of Domain-Specific AGI being demonstrated, regardless of the safety implications.

Multi-Domain AGI (October 2027)

Assumption 1: Recursive self-improvement generalizes beyond software. Once a system gets better at getting better, the capability gains must transfer across domains. This is not guaranteed. A system that optimizes its own code generation may produce no improvement in scientific reasoning or legal analysis if those capabilities depend on different architectural properties. Our model assumes that the meta-capability of self-improvement is domain-general, but this has never been demonstrated.

Assumption 2: Non-software evaluation infrastructure can be built fast enough. Verifying that an AI system has achieved autonomous performance in scientific research or professional services requires evaluation frameworks that do not yet exist. Software has test suites and deployment metrics. Drug discovery has clinical trials that take years. Our prediction assumes that evaluation infrastructure for non-software domains can be built or adapted within months, not years. If verification is the bottleneck, capability may exist long before we can confirm it.

8.2 Leading Indicators

For each prediction, we specify concrete evidence that should be visible approximately 12 months before the point estimate. These are not predictions themselves but tripwires: if we do not observe them by the specified date, the corresponding prediction is in trouble.

Domain-Specific AGI (watch by March 2026)

These indicators should be observable now or within weeks:

- SWE-bench "hard" scores exceed 80% (current state-of-the-art is approximately 50–60%).
- At least one autonomous agent publicly ships a complete application from a single specification prompt with no human intervention at any stage.
- An autonomous agent successfully manages a production deployment for one or more weeks without human intervention, including incident response.
- Error recovery rate on multi-step tasks improves from the current approximately 60% to 85% or higher.

Recursive Self-Improvement (watch by July 2026)

- A major lab publishes results showing AI-assisted architecture search producing measurable training efficiency gains.
- AlphaEvolve or a similar system demonstrates capability improvements that transfer beyond the narrow optimization target.
- The open-source community replicates AI-assisted AI improvement with documented methodology and reproducible results.

Multi-Domain AGI (watch by October 2026)

- AI systems demonstrate autonomous research capability, including designing and executing experiments rather than only generating text about them.
- Cross-domain transfer learning shows step-function improvement: a model trained primarily for code excels at novel scientific tasks without domain-specific fine-tuning.
- At least one automated laboratory (consistent with Hassabis's stated 2026 plans) produces a peer-reviewable result.

8.3 Falsification Triggers

The following observations would cause us to push the corresponding prediction at least two years later. These are not minor setbacks but structural failures of our causal model.

Domain-Specific AGI

- SWE-bench hard scores plateau below 70% through the end of 2026, indicating that the current improvement trajectory has stalled.
- No autonomous agent completes even a Level 2 project (let alone Level 3) by Q3 2026.
- A major compute supply disruption—export controls, energy constraints, or supply chain failures—slows frontier model training by six or more months.

Recursive Self-Improvement

- Domain-Specific AGI does not arrive by September 2027 (the outer bound of our confidence window), breaking the causal chain.
- Labs publish results showing that AI-assisted AI improvement produces diminishing or negative returns, suggesting that self-improvement is harder than "just another software project."
- Regulatory action explicitly prohibits autonomous AI self-modification at frontier labs, removing the competitive pressure that our model assumes drives adoption.

Multi-Domain AGI

- Recursive self-improvement does not produce cross-domain gains: improvements remain narrow to the domain being optimized, falsifying the generalization assumption.
- Scientific peer review proves systematically resistant to AI-generated research, regardless of quality, creating a verification bottleneck that cannot be resolved quickly.

- Evaluation infrastructure for non-software domains takes years to build rather than months, meaning capability may exist but cannot be confirmed within our timeline.

8.4 Addressing Known Weaknesses

Scaling and Diminishing Returns

The 2025 correction in market sentiment reflected genuine evidence that pure parameter scaling yields diminishing capability returns at the frontier. Our model does not assume scaling alone solves the remaining problems. It assumes that architectural innovation—better planning algorithms, more robust error recovery, longer effective memory—combines with scaling to close the gap. The editor signal weights the observed improvement rate from daily tool use over the market correction, which may have overcorrected in response to premature product launches rather than a genuine capability plateau. If scaling hits a hard wall and architectural innovation does not compensate, the timeline extends. Architectural innovation has a less predictable trajectory than scaling because it depends on discrete breakthroughs rather than continuous investment.

Data and Learning Limits

Web-scale text data is largely exhausted as a training signal. The next generation of training data comes from three sources: (1) interaction data from millions of agent deployments, capturing real-world usage patterns and failure modes; (2) synthetic data from verified code execution, where correctness is mechanically verifiable because code either runs or it does not; and (3) reinforcement learning from real-world deployment outcomes, including uptime metrics, user satisfaction, and incident resolution. Software engineering is uniquely advantaged here because the feedback signal is concrete and automated. Domains where correctness requires human judgment (legal analysis, scientific peer review) will lag.

Reliability Threshold

Our definition of "operate" implicitly requires high reliability. A system that deploys a platform for 1,000 users but crashes constantly, leaks data, or requires daily human intervention would not meet our criteria. Specifically, "operate" requires: maintaining greater than 99% uptime over the 30-day evaluation period, responding to incidents within reasonable SLAs, and handling edge cases without human escalation for routine issues. This reliability requirement is a meaningful additional hurdle beyond raw capability.

Hardware, Energy, and Economics

Current frontier model training runs cost $100M–$1B. Inference costs are dropping at roughly 10x per year. Our predictions assume three economic conditions: (1) training costs plateau or decrease through efficiency gains and architectural innovation; (2) inference becomes cheap enough for continuous autonomous agent operation, in the range of $1–10 per hour; and (3) no major compute supply shock disrupts availability. If inference costs for autonomous agents remain above $100 per hour, the deployment timeline extends substantially even if the underlying capability exists, because the 1,000-user requirement implies economically viable operation.

Reference Class Validity

The PDM was validated on historical physical technologies: internet adoption, mobile computing, cloud infrastructure. Software capability development may follow fundamentally different dynamics, with faster iteration cycles and lower capital requirements. This is a known limitation. However, the PDM contributes one signal out of five in our ensemble, not a sole prediction. The blind holdout validation ($p < 0.0001$) suggests the framework captures something real about precondition convergence patterns, even if the exact timeline calibration may differ for software-native capabilities.

Capability vs. Deployment Gap

Our 1,000-user requirement bridges the capability-deployment gap by definition: it requires not just laboratory demonstration but actual production use. However, we acknowledge that safety testing, liability frameworks, and regulatory approval could add 6–18 months between "it works in the lab" and "1,000 people are relying on it." This gap is partially captured in our ±6 month confidence windows, but if regulatory response is aggressive, the confidence windows may prove too narrow.

9 Predictions

Table 1 presents our point estimates and confidence windows. Point estimates are derived by identifying the date at which the ensemble cumulative probability crosses 25%—roughly the median of the "more likely than not within the window" threshold—then adjusting for sequential dependency between milestones. These represent the ensemble model output, not the editor signal alone.

Table 1: AGI timeline predictions — point estimates with ±6-month confidence windows.

Prediction	**Point Estimate**	**Confidence Window**
Domain-Specific AGI (SWE)	March 2027	Sept 2026 – Sept 2027
Recursive Self-Improvement	July 2027	Jan 2027 – Jan 2028
Multi-Domain AGI	October 2027	Apr 2027 – Apr 2028

The predicted sequence is not arbitrary. Software engineering falls first because it has the tightest feedback loop of any domain: code either works or it does not, millions of developers provide training signal, and outputs are self-verifying through tests and deployment. This makes it the domain where AI capability improvement translates most directly into autonomous performance.

Recursive self-improvement follows because AI infrastructure is software. A system capable of building and operating production software platforms can be directed at its own training pipeline, inference stack, and architecture. The four-month gap reflects the additional challenge of self-referential optimization over general software engineering.

Multi-domain AGI arrives last, accelerated by recursive improvement. Once a system is getting better at getting better, capability gains in non-software domains follow as the system optimizes its own performance in those areas. The three-month gap between recursive improvement and multi-domain AGI is our most aggressive prediction, reflecting the view that recursive improvement creates a rapid capability cascade.

The full sequence spans approximately seven months. This is a bold claim, and we are aware that compressed timelines are the most common failure mode in technology forecasting.

9.1 Ensemble vs. Editor Divergence

The editor signal is 3–4x more bullish than the model average across all three predictions for the 2027–2028 window. Three factors drive this divergence:

1. **Swarm feedback.** Millions of developers using AI coding tools generate an unprecedented volume of implicit training signal. This feedback density has no historical parallel and is not captured by the PDM's historical-analogy approach.
2. **Compound improvement curves.** Each capability gain makes the next gain easier. The editor observes this compounding daily; the other signals model it poorly because they treat improvement as approximately linear.
3. **Practitioner weighting.** The editor weights the practitioner experience more heavily than market sentiment or expert opinion, on the grounds that daily users of a technology are better calibrated on its trajectory than observers.

We report both values transparently and let readers weight them according to their own epistemics.

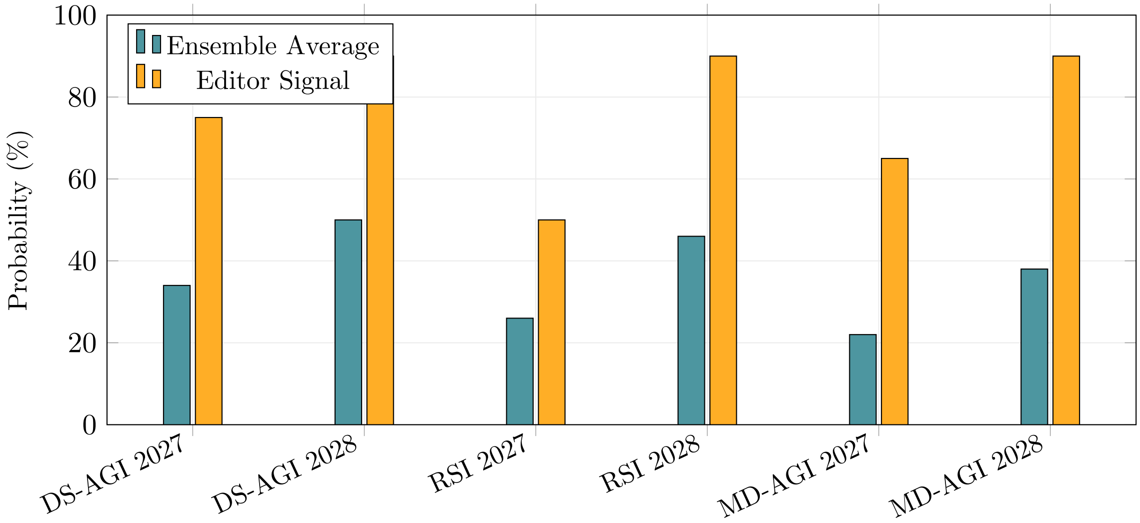

Figure 4: Ensemble average vs. editor signal for each prediction at the 2027 and 2028 horizons. The editor signal is consistently 2–4x more bullish, reflecting practitioner experience with compound improvement rates.

10 Ensemble Data Tables

Tables 2–4 present the full five-signal data for each prediction.

Table 2: Domain-Specific AGI (Software Engineering) — five-signal ensemble probabilities.

Timeframe	PDM	Markets	Experts	LLM	Editor	Ensemble
By end 2026	10%	∼8%	∼12%	4%	7.5%	∼8%
By end 2027	25%	∼20%	∼30%	18%	75%	∼34%
By 2028	45%	∼35%	∼45%	33%	90%	∼50%
By 2030	70%	∼60%	∼65%	60%	95%	∼70%

Table 3: Recursive Self-Improvement — five-signal ensemble probabilities.

Timeframe	PDM	Markets	Experts	LLM	Editor	Ensemble
By end 2026	5%	N/A	∼10%	8%	5%	∼7%
By end 2027	15%	N/A	∼20%	20%	50%	∼26%
By 2028	30%	N/A	∼30%	35%	90%	∼46%
By 2030	55%	N/A	∼50%	60%	95%	∼65%

Table 4: Multi-Domain AGI — five-signal ensemble probabilities.

Timeframe	PDM	Markets	Experts	LLM	Editor	Ensemble
By end 2026	2%	∼3%	∼3%	1%	5%	∼3%
By end 2027	10%	∼12%	∼15%	8%	65%	∼22%
By 2028	25%	∼25%	∼30%	20%	90%	∼38%
By 2030	50%	∼45%	∼50%	45%	95%	∼57%
By 2032	70%	∼60%	∼65%	60%	99%	∼71%

11 Verification and Registration

All three predictions are registered on public forecasting platforms:

- **Metaculus:** We will create or link to existing questions matching our definitions as closely as possible, with clarifying comments where our criteria differ from the platform's resolution criteria.
- **Manifold Markets:** Binary markets created for each prediction with explicit resolution criteria matching Section 2.
- **Future Shock:** Continuous monitoring and quarterly updates published at `future-shock.ai`, including any revisions to our estimates with full justification.

Predictions will be graded on calibration (did the probabilities match observed frequencies?), resolution (did the point estimates fall within the confidence windows?), and reasoning quality (were the stated assumptions and mechanisms validated or falsified?).

We commit to publishing a detailed post-mortem for each prediction at resolution, regardless of outcome. Getting predictions right is valuable; understanding why predictions fail is more valuable.

12 Limitations

This framework has significant limitations that readers should weigh when interpreting our predictions.

PDM validation gap. The Precondition Density Model was validated on historical technology convergence events (internet, mobile, cloud). AI development may follow fundamentally different dynamics. The model has no track record on predicting AI milestones specifically.

Market immaturity. Prediction markets on AGI-class questions are thin, attract ideological participants, and have no resolved precedents to calibrate against. Market probabilities may reflect sentiment rather than information.

Expert bias. The expert positions we compile are overwhelmingly from individuals with commercial interests in AI development. Optimistic timelines attract funding; pessimistic timelines do not. We adjust for this bias but cannot eliminate it.

Editor signal limitations. The editor signal is n=1 by construction. It captures practitioner experience that other signals miss, but it is not replicable, not blinded, and subject to all the cognitive biases that affect individual judgment.

Small calibration sample. Future Shock's prediction engine has a limited number of resolved predictions available for calibration. Our confidence in our own calibration is necessarily low at this stage.

Regulatory risk. Government intervention—compute thresholds under the EU AI Act, US executive orders, or export controls on training hardware—could materially delay timelines. Our model does not incorporate regulatory scenarios.

Safety implications of RSI. Recursive Self-Improvement, if achieved, raises significant AI safety concerns that are beyond the scope of this paper but should inform any policy response to these predictions. Publishing aggressive RSI timelines may also influence competitive dynamics between labs.

Definition sensitivity. Small changes to our achievement criteria (e.g., requiring 10,000 users instead of 1,000, or 10% benchmark improvement instead of 5%) could shift predictions by years. We have chosen thresholds that we believe are meaningful, but reasonable people will disagree.

13 Conclusion

We are making three specific, falsifiable claims: Domain-Specific AGI in software engineering by March 2027, Recursive Self-Improvement by July 2027, and Multi-Domain AGI by October 2027. The predicted sequence matters as much as the dates. Software engineering falls first because code is self-verifying and developer feedback is abundant. That capability, turned inward, produces recursive self-improvement. Recursive improvement accelerates everything else.

The full sequence spans seven months. If this timeline holds, 2027 will be the most consequential year in the history of computing. If it does not, our prediction engine takes a credibility hit that we will document and learn from publicly.

We publish these predictions not because we are certain, but because the field needs more specificity and less hand-waving. Vague predictions are unfalsifiable, and unfalsifiable predictions are worthless. Whether we are right or wrong, the act of being precise creates value: it forces us to articulate our assumptions, exposes our reasoning to scrutiny, and gives the forecasting community concrete claims to update on as evidence accumulates.

The ensemble says 2027. The editor says sooner. We will find out together.

References

[1] S. Altman, "The Intelligence Age," OpenAI Blog, September 2024.

[2] D. Amodei, "Machines of Loving Grace," Dario Amodei's Blog, October 2024.

[3] M. Morris et al., "Levels of AGI: Operationalizing Progress on the Path to AGI," Google DeepMind, arXiv:2311.02462, 2023.

[4] Metaculus, "When will the first general AI system be devised, tested, and publicly announced?" `https://www.metaculus.com/questions/5121/`.

[5] T. Trinh et al., "AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery," Google DeepMind, 2025.

[6] Y. LeCun, "A Path Towards Autonomous Machine Intelligence," OpenReview, 2022.

[7] G. Marcus, "Deep Learning Is Hitting a Wall," Nautilus, March 2022.

[8] G. Hinton, interview with CBS *60 Minutes*, October 2023.

[9] E. Musk, remarks at the 2024 Milken Institute Global Conference, May 2024.

[10] I. Sutskever, "Sequence to Sequence Learning with Neural Networks" and subsequent public statements on AGI timelines, 2024–2025.