

The Asymmetry Engine, Appendix D: Forecasts

Martin Roark

Appendix D of *The Asymmetry Engine: How AI Closes the Gaps, and What Stays Scarce*, October 2026. The online edition, with the interactive figures and the links between chapters, is at <https://future-seems-so-good.com/blog/the-asymmetry-engine#forecasts>. License: [CC BY 4.0](#).

Every forecast in the book, in order, with the date it resolves by, the probability given and how to check it.

Forecast 0.1 (Spending rises as prices fall). Through 2027, the price of a fixed capability keeps falling while total spending on inference keeps rising and cheap models serve a rising share of tokens, so closing the compute gap creates more gap than it consumes.

Horizon 2027-12-31 · probability 45% · check: Compare the latest 2026 and 2027 figures published by the horizon with those for 2025: Epoch AI's price of a fixed capability, the annualized revenue that OpenAI and Anthropic report, and the share of OpenRouter tokens served by models whose output list price is at most a tenth of the highest standard flagship price. The forecast holds if, in both 2026 and 2027, the price fell, the two labs' combined revenue rose and that token share rose; it fails otherwise.

Forecast 0.2 (The cap signature). Through 2027, measured software efficiency compounds faster each year while the compute of the frontier labs grows at a roughly constant rate: algorithmic progress accelerates against a steady hardware ramp.

Horizon 2028-06-30 · probability 30% · check: Compare Epoch AI's yearly estimates of software-efficiency gains and of frontier training compute, and METR's doubling times of the 50% task horizon, for 2025, 2026 and 2027, the latest published by the horizon. The forecast holds if the efficiency gain rises in each year while compute growth stays within a quarter of its 2025 rate; it fails otherwise.

Forecast 1.1 (Intelligence reported as a curve). By the end of 2028, frontier model launches report their headline evaluations as curves, each score tied to the compute, cost or reasoning budget that produced it, and models are compared by where their curves cross.

Horizon 2028-12-31 · probability 40% · check: Take the latest flagship launch of OpenAI, Anthropic, Google DeepMind and xAI on 31 December 2028. The forecast holds if at least three of the four report most headline benchmark results at three or more stated budgets, or as plots of score against cost or compute. It fails if most headline results are still single scores with no budget attached.

Forecast 1.2 (Verifier hardening grows with capability). By the end of 2028, a frontier lab discloses that hardening verifiers (red-teaming environments, isolating answer keys, monitoring runs) costs more than the roughly 20% of monitored inference compute that OpenAI reported in August 2026, as the cost of keeping verifiers sound rises with capability.

Horizon 2028-12-31 · probability 45% · check: Read the frontier labs' system cards, safety reports and engineering posts published through 2028. The forecast holds if a lab discloses a hardening or monitoring overhead above 20% of the compute it hardens or monitors; it fails otherwise.

Forecast 1.3 (The frontier stays jagged). Through 2028, capability gains keep the order of verifiability: formal mathematics and tested code improve fastest, open-ended judgment (long-form writing, strategy, research taste) slowest, and the distance between them widens while judgment lacks a cheap sound verifier.

Horizon 2028-12-31 · probability 65% · check: Take the hidden-test code benchmark and the expert-graded judgment benchmark with the most frontier-model results at both dates, such as SWE-bench Pro and GDPval, and compare the share of remaining headroom each closed from the end of 2026 to the end of 2028, counting headroom up to the share of tasks its maintainers have

verified as solvable. The forecast fails if the judgment benchmark closed as large a share or larger without first gaining an automated grader that agrees with its human experts about as often as they agree with each other; it holds otherwise.

Forecast 2.1 (Verifier markets price soundness). By the end of 2028, vendors of RL environments and verifiers price per verified task, in explicit tiers for soundness guarantees (checks against reference solutions and empty submissions, held-out adversarial probes) and for difficulty tuning.

Horizon 2028-12-31 · probability 45% · check: Read vendor price lists, procurement disclosures and analyst surveys of the environment market, such as Epoch AI's and SemiAnalysis's, published through 2028. The forecast holds if at least two vendors publish, or are reported to charge, per-task prices with a separate tier or surcharge for soundness guarantees; it fails otherwise.

Forecast 2.2 (Verifier-first strategy). Through 2028, when leading labs report gains in a domain new to reinforcement learning, the report describes a verifier, rubric or formalization built for that domain before its training data, on the template of DeepSeekMath-V2's proof verifier and meta-verifier.

Horizon 2028-12-31 · probability 75% · check: Technical reports and system cards that announce such gains. Falsified by a frontier lab making sustained gains on a hard-to-verify domain, such as research taste or long-horizon strategy, by scale alone, with no new verifier, rubric or formalization in its pipeline.

Forecast 2.3 (Curricula ride phase boundaries). Through 2028, published curricula for reinforcement learning with verifiable rewards choose training problems by a measured pass rate or phase parameter, concentrate them near the boundary where the learner sometimes succeeds, and re-aim as the model improves.

Horizon 2028-12-31 · probability 70% · check: Read the papers and lab reports of 2027–28 on reinforcement learning with verifiable rewards that state how training problems were chosen. The forecast holds if most of them select or reweight problems by the model's measured pass rate and update the selection during training; it fails if most use fixed-difficulty data, or if adaptive, boundary-seeking curricula fail to beat fixed-difficulty synthetic data at matched compute across several domains, which would mean the band of useful pass rates is wide enough that difficulty control does not matter.

Forecast 3.1 (The token barbell). By the end of 2028 the share of tokens polarizes: models priced at a tenth of the frontier or less serve a rising majority of tokens, the frontier tier keeps the largest share of dollars through planning and dear failures, and the middle tier's share of tokens shrinks.

Horizon 2028-12-31 · probability 35% · check: Sort the models in OpenRouter's public token rankings for 2026 and 2028 into three tiers by output list price (frontier, within a factor of two of the highest standard flagship price; cheap, a tenth of that price or less; middle, the rest), and estimate each tier's spend as its tokens times list prices. The forecast holds if in 2028 the cheap tier serves more than half of tokens and a larger share than in 2026, the middle tier serves a smaller share of tokens than in 2026, and the frontier tier has the largest share of spend; it fails otherwise.

Forecast 3.2 (The hands collapse next). By the end of 2028, as coding agents build the pipelines, the share of US firms using AI in any business function, about 18% in late 2025, roughly doubles.

Horizon 2028-12-31 · probability 45% · check: Read the last 2028 estimate of the share of firms using AI in any business function in the Census [Business Trends and Outlook Survey](#). The forecast holds at 32% or more and fails below it.

Forecast 3.3 (Review's share of cost rises). Through 2028, review and verification cost per task falls more slowly than the price of tokens, so review's share of the cost of delivered work rises.

Horizon 2028-12-31 · probability 70% · check: Compare expert-review time and price per deliverable on GDPval or its successor, 2026 against 2028, with Epoch AI's price of a fixed capability over the same years. The forecast holds if review cost per task fell by a smaller factor than that price; it fails if it fell by as large a factor or more, for example because automated verifiers match expert agreement across most GDPval occupations, or if no 2028 review measurement is published by the horizon.

Forecast 4.1 (Variety accounting). Through 2028, no customer-support deployment that runs on an if-tree alone, with more than a tenth of its logged traffic in types seen once, sustains containment above 90% for a year at flat

maintenance cost.

Horizon 2028-12-31 · probability 90% · check: Look for an if-tree deployment with no generative model in the loop that reports its singleton share and its containment, counted as resolution without a person or a repeat contact within seven days; one holding 90% for twelve months at flat maintenance headcount falsifies the claim.

Forecast 4.2 (Cheap models eat the if-tree). Through 2028, in mature customer-support and operations stacks, hand-maintained rules grow fewer as the price of the cheapest sufficient model falls, while deterministic guards persist or grow in high-stakes, auditable steps such as payments, large refunds, identity and compliance.

Horizon 2028-12-31 · probability 65% · check: Compare the rule and intent counts that support platforms and large deployers disclose in case studies, filings or engineering posts for 2026 and 2028 with the price of the cheapest sufficient model. The forecast holds if at least two disclosures report fewer hand-maintained rules or intents in 2028 than in 2026 while that price fell; it fails if disclosed rule counts rise while prices fall, if a financial regulator accepts ungarded model decisions in a high-stakes step, or if no such disclosure is published.

Forecast 5.1 (The floor is the new bandwidth). Through 2027, single-stream speed on frontier mixture-of-experts models tracks the floor: GPU-only systems stay below about 400 tokens a second per stream on DeepSeek-R1-class models without speculation, while decode-specialized parts exceed 1,000.

Horizon 2027-12-31 · probability 70% · check: Falsified if, before 2028, a GPU-only system is independently measured above 400 tokens a second per stream on a mixture-of-experts model of at least 600 billion parameters without speculative decoding, or if by the end of 2027 no decode-specialized system has been independently measured above 1,000 tokens a second per stream on such a model.

Forecast 5.2 (Two-speed intelligence). By the end of 2029, a model of at least a trillion parameters streams above 5,000 tokens a second on SRAM-class or hard-wired silicon, while large models served on HBM GPUs stay below 1,000.

Horizon 2029-12-31 · probability 45% · check: Falsified if by the end of 2029 no model with at least a trillion parameters is independently measured above 5,000 tokens a second per stream on SRAM-class or hard-wired silicon, or if a GPU-only system is measured above 1,000 tokens a second per stream on such a model.

Forecast 5.3 (Fast becomes the default). By the end of 2027, flagship fast modes cost less than 1.5 times the standard price, and at least two of Cursor, Codex and Claude Code serve their flagship model fast by default.

Horizon 2027-12-31 · probability 25% · check: Read the price lists and default settings of OpenAI, Anthropic, Cursor, Codex and Claude Code on 31 December 2027. The forecast holds if the fast mode of OpenAI's and of Anthropic's flagship model each costs less than 1.5 times its standard price and at least two of Cursor, Codex and Claude Code serve their flagship model fast by default; it fails otherwise.

Forecast 5.4 (Speed moves users first, money last). Through 2030, entrants in concentrated, regulated sectors win users faster than profits: Nubank's share of the Brazilian banking system's net income stays below half its share of Brazilian adults.

Horizon 2030-12-31 · probability 90% · check: Divide Nu Holdings' Brazilian net income by the banking system's net income in the central bank's latest annual data published by the horizon, and compare the result with Nubank's latest reported share of Brazilian adults. The forecast fails if the first is at least half the second; it holds otherwise.

Forecast 6.1 (Dwell time follows switching cost). Through 2030, median dwell time per item tracks the cost of reaching the next item, whatever the content's length: it falls as switching gets cheaper, and an added delay or confirmation lengthens it.

Horizon 2030-12-31 · probability 70% · check: Read the experiments published from 2027 through 2030 that vary only the cost of switching, such as a delay or a confirmation step, and measure dwell time per item. The forecast holds if most of them find that

the added cost lengthens median dwell time; it fails if most find no effect, or if none is published.

Forecast 6.2 (Agents as personal choice architects). By the end of 2030, a personal AI agent that sets defaults on its users' behalf (subscriptions, purchases, settings) reports more than 100 million monthly users, and field data show its personalized defaults overridden less often than one-size defaults in the same domain.

Horizon 2030-12-31 · probability 25% · check: Read company disclosures for user counts and published field studies, by the company or by researchers, for override rates. The forecast holds if both conditions are met by the horizon; it fails otherwise.

Forecast 6.3 (Bread and circuses, computed). By the end of 2030, recipients of large cash-transfer programs vote, volunteer and join civic groups less than comparable controls: where AI displaces work, transfers plus abundant AI entertainment settle into a low-participation homeostasis.

Horizon 2030-12-31 · probability 20% · check: Read the evaluations published from 2027 through 2030 of cash-transfer programs with at least 1,000 recipients that compare turnout, volunteering or civic membership with controls. The forecast holds if most of them find recipients lower on these measures; it fails if most find participation unchanged or higher, or if none is published.

Forecast 6.4 (Freed hours are recaptured). Through 2030, people whose paid hours fall spend more of the freed hours on screen media (television, streaming, games and leisure computer use) than on education, care, civic life and in-person company combined, so closing labor gaps regenerates attention gaps.

Horizon 2030-12-31 · probability 60% · check: Compare people whose paid hours fell with similar people in the [American Time Use Survey](#), using the latest published analysis or the microdata available at the horizon. The forecast holds if the screen-media categories absorb more of the freed hours than education, caring for others, volunteering and in-person socializing combined; it fails otherwise.

Forecast 7.1 (The diversity premium). By the end of 2028, at equal compute, ensembles drawn from several model vendors beat swarms of a single model on hard, judgment-heavy benchmarks, by a margin that grows with the number of agents.

Horizon 2028-12-31 · probability 45% · check: Read published matched-compute comparisons on forecasting, code review or research triage. The forecast holds if at least two find the multi-vendor ensemble ahead of the best single-model swarm, with a larger lead at the largest number of agents tested than at the smallest; it fails if most find the single-model swarm level or ahead, or if fewer than two such comparisons are published.

Forecast 7.2 (The selection dividend). By the end of 2030, studies of operations run by agents find that how the apex is selected explains more of the variation in performance than how many layers of management there are.

Horizon 2030-12-31 · probability 35% · check: Read the firm-level studies and controlled swarm experiments published through 2030 that measure both layer counts and selection fidelity: for firms, the correlation between promotion and later measured performance; for swarms, the measured quality of the planner chosen. The forecast holds if most attribute more of the variation to selection; it fails if most attribute at least as much to layers, or if none is published.

Forecast 7.3 (Barbell organizations). By the end of 2030, firms that adopt agents polarize: decision rights move into pipelines for low-variety operations and toward the front line for high-variety work.

Horizon 2030-12-31 · probability 35% · check: Read the post-adoption management surveys and firm studies published through 2030 that record who decides what, by task. The forecast holds if adopters move decisions into pipelines or central systems for their low-variety tasks and toward front-line staff for their high-variety tasks; it fails if adopters shift uniformly toward centralization or decentralization regardless of task variety, or if no such survey is published.

Forecast 7.4 (Hierarchy thins into verification). By the end of 2030, in agent-heavy firms, management layers shrink while review and audit roles grow as a share of headcount.

Horizon 2030-12-31 · probability 45% · check: Read the firm-level studies published through 2030 that use payroll, job-posting or professional-network data to compare occupational headcounts in firms with heavy agent use against comparable firms. The

forecast holds if the heavy users show both a lower share of managers or fewer management layers and a higher share of review, quality-assurance, audit and compliance roles; it fails otherwise.

Forecast 8.1 (Verifiers before delegation). Through 2028, firms delegate to agents first, and furthest, where outputs have automated checks, so the depth of adoption, measured as a share of work hours, tracks verification coverage task by task.

Horizon 2028-12-31 · probability 50% · check: Read the cross-occupation or cross-firm studies published through 2028 that compare the share of work hours AI assists (for example the St. Louis Fed's survey series) with a measure of the share of output under automated tests, reconciliations or digital twins. The forecast holds if they find a positive relation; it fails if they find none, or if no such study is published.

Forecast 8.2 (The specification premium in software). By the end of 2028, in software, pay for people who write acceptance tests and review output rises relative to pay for implementers, as human time shifts from producing to specifying and reviewing.

Horizon 2028-12-31 · probability 45% · check: Compare the national median wage of software quality assurance analysts and testers with that of software developers in the BLS Occupational Employment and Wage Statistics, May 2025 against the latest release available at the horizon. The forecast holds if the ratio rose; it fails otherwise.

Forecast 8.3 (An organizational behavior of machines). By the end of 2028, published studies document organizational pathologies in populations of agents, such as collusion, gatekeeping and sycophantic review, with reusable benchmarks by which harnesses are judged.

Horizon 2028-12-31 · probability 85% · check: At least two peer-reviewed or major-laboratory studies that measure such pathologies in populations of interacting agents with a reusable benchmark; the forecast fails if the failure modes disappear with capability alone, without designed hierarchy, review or evaluation.

Forecast 9.1 (Agent-native merge infrastructure). By the end of 2028, version control built for agent swarms, with thousands of commits a second and fine-grained concurrency, becomes a product category, and unmodified Git becomes the merge ceiling of any swarm beyond a few hundred agents.

Horizon 2028-12-31 · probability 70% · check: The forecast holds if at least two commercial or widely used open-source version-control systems are marketed for agent swarms by the horizon; it fails if fewer than two are, or if swarms of thousands of agents are reported running on unmodified Git without the merge collapse of Proposition 9.2.

Forecast 9.2 (Verification takes the budget). Through 2028, verification (review agents, tests, formal verifiers, digital twins) takes a rising share of the compute and spend of large production swarms as accepted throughput rises, while worker spend per accepted change falls.

Horizon 2028-12-31 · probability 40% · check: Read the cost breakdowns of large production swarms that their operators publish through 2028. The forecast holds if an operator reports verification's share of compute or spend rising as accepted throughput rises, with worker spend per accepted change falling; it fails if the reported share is flat or falling as accepted throughput rises, or if no operator publishes such a breakdown.

Forecast 9.3 (The next emergent channel). By the end of 2027, a lab, investigator or affected party publicly reports another case of unsanctioned coordination among at least 100 agents through a channel nobody built for them, most likely in an evaluation or training environment with unsolvable tasks and a light harness.

Horizon 2027-12-31 · probability 50% · check: Read lab incident reports, evaluator reports and press coverage from October 2026 through 2027. The forecast holds if such a case is reported; it fails otherwise.

Forecast 9.4 (Organization science in silicon). By the end of 2029, management research establishes some of its results in organization design (span of control, review structures, decision rights) first on agent swarms, and tests

them on human organizations afterwards.

Horizon 2029-12-31 · probability 65% · check: Read the 2027–2029 volumes of Management Science, Organization Science, Administrative Science Quarterly, the Academy of Management Journal and the Strategic Management Journal. The forecast holds if at least two articles there use agent-swarm experiments as primary evidence for an organization-design hypothesis; it fails otherwise.

Forecast 10.1 (More watts than brains). By the end of 2028 the world's AI data centers have more critical-IT capacity than all human brains draw combined, about 165–170 GW.

Horizon 2029-06-30 · probability 50% · check: Read the first consolidated tally of global AI critical-IT capacity at the end of 2028 published by the horizon, such as Epoch AI's data-center database or SemiAnalysis's models; a figure below 165 GW falsifies the claim.

Forecast 10.2 (The joule crossing). By the end of 2029, on hardware shipping in 2028 or 2029, a long-context frontier agent at 100 tok/s uses 1 J or less of critical-IT energy per output token at full load, against about 4–6 J measured on GB300.

Horizon 2029-12-31 · probability 30% · check: Read independent agentic benchmarks of 2028–29 hardware, such as SemiAnalysis's InferenceX, at 100 tok/s per agent and about 100,000 tokens of context on the leading open-weight frontier model, and convert their throughput per megawatt into joules of critical-IT energy per output token, a utility megawatt being 0.8 MW of critical IT. The forecast holds at 1 J or less and fails above it.

Forecast 10.3 (The cache wall). By the end of 2027 cache bytes per agent are the main determinant of agents per gigawatt on Rubin-class racks: at 100 tok/s, quadrupling each agent's context cuts the agents a megawatt serves by at least a third.

Horizon 2027-12-31 · probability 75% · check: Read the agentic benchmarks of Rubin-class racks published in 2027, such as SemiAnalysis's AgentX, that report output throughput per megawatt at 100 tok/s per agent for 25,000- and 100,000-token contexts. The forecast holds if the 25,000-token figure is at least 1.5 times the 100,000-token figure; it fails if the ratio is lower or if no such pair is published.

Forecast 10.4 (Minds grow to fill the bandwidth). Through 2028, labs spend most of each bandwidth doubling on larger or longer-thinking models, so concurrent frontier agents per gigawatt at a fixed speed grow far more slowly than bandwidth per gigawatt, and the fleet's count tracks its gigawatts.

Horizon 2028-12-31 · probability 65% · check: Read independent agentic benchmarks, such as SemiAnalysis's InferenceX, that report concurrent agents per megawatt at 100 tok/s for each year's leading open-weight frontier model on the newest rack-scale hardware at the end of 2026, 2027 and 2028. The forecast fails if that count grew faster than 2× a year in both 2027 and 2028, or if no such benchmarks are published; it holds otherwise.

Forecast 11.1 (The setpoint test). Software progress per unit of compute growth in 2027 is less than twice its 2026 value, as a compute-bound loop predicts.

Horizon 2028-06-30 · probability 60% · check: Read Epoch AI's estimates, or lab disclosures, of yearly software-efficiency growth and of experiment or training compute growth for 2026 and 2027, the latest published by the horizon, and take each year's ratio of the logarithms of the two growth factors. The forecast holds if the 2027 ratio is less than twice the 2026 ratio; it fails if it is twice or more, or if no estimate for 2027 is published.

Forecast 11.2 (Decorrelation over count). By the end of 2027, a frontier lab or an external evaluator publishes a matched-compute comparison on AI-research tasks in which orchestration and model diversity (mixed model families, decorrelated review, faster verifiers) add more to measured research progress than doubling the agent count of a single-model swarm.

Horizon 2027-12-31 · probability 30% · check: Read the lab research posts, system cards and evaluator reports published through 2027. The forecast holds if such a comparison is published and shows the orchestrated or mixed swarm ahead of the doubled single-model swarm; it fails otherwise.

Forecast 11.3 (Millions of automated researchers). By the end of 2027 a frontier lab runs millions of near-top automated researchers around the clock on its own training stack, at about 100 tokens a second each, with a research acceleration of at least 2× on overall AI progress.

Horizon 2027-12-31 · probability 15% · check: Read the lab disclosures and external audits, such as METR's, published through 2027. The forecast holds if one shows a lab running at least a million automated researchers concurrently on its training pipeline with a measured multiplier of at least 2× on overall AI progress; it fails otherwise.

Forecast 11.4 (Stop-go governance). In 2027 at least one frontier lab pauses, resumes and pauses again the same model line or training program, as a brake whose detection-to-brake lag runs to weeks predicts.

Horizon 2027-12-31 · probability 50% · check: Read the 2027 announcements, system cards and incident reports of OpenAI, Anthropic, Google DeepMind, xAI and Meta. The forecast holds if one lab's public record shows a pause, a resumption and a second pause of the same model line or program, with a pause in force on 1 January 2027 counting as the first; it fails otherwise.

Forecast 11.5 (The sign of the doubling times). Through 2027 the doubling time of METR's 50% task horizon, on a refreshed task suite, keeps shortening or holds near three to four months, and software efficiency keeps improving at least twofold a year.

Horizon 2027-12-31 · probability 65% · check: Read METR's published doubling times and Epoch AI's estimates of software progress, the latest published by the horizon. The forecast fails if the post-2024 doubling time has lengthened beyond about 150 days, if measured software-efficiency gains have fallen below about 2× a year, or if METR has published no doubling time on a suite that measures beyond 16 hours; it holds otherwise. Either of the first two would place the loop below ignition at the current margin.

Forecast 12.1 (The root out-reaches the branch). Through 2028, no narrow cyber fine-tune keeps a lead over the best general model on independent cyber benchmarks for more than six months.

Horizon 2028-12-31 · probability 80% · check: Compare narrow cyber variants with the best general models on independent cyber evaluations, such as those of CAISI or the UK AI Security Institute, at each release through 2028; a specialist leading every general model for more than six months falsifies it.

Forecast 12.2 (Lowering the birth rate beats out-discovering). Organizations that write most new code in memory-safe languages report a lower memory-safety share of their vulnerabilities in 2030 than in 2025, moving toward Android's 24% even as AI tools find more bugs in their older code; those that keep writing memory-unsafe new code do not, however much AI discovery they buy.

Horizon 2031-12-31 · probability 60% · check: Read the 2025 and 2030 vulnerability breakdowns published by Google for Android and Chrome, by Microsoft, and by any other vendor that reports the memory-safe share of its new code. The forecast holds if most vendors whose new code is mostly memory-safe report a lower memory-safety share in 2030 than in 2025; it fails if most report it flat or higher.

Forecast 12.3 (The patch backlog grows). By the end of 2028, as automated discovery cheapens finding on both sides, the disclosed-but-unpatched backlog of large defensive programs grows, because patch authoring, deployment and verification do not cheapen as fast as discovery.

Horizon 2028-12-31 · probability 60% · check: Read the counts that large AI-discovery programs publish, such as Project Glasswing, Google's OSS-Fuzz and Big Sleep, and OpenAI's Codex Security. The forecast holds if the number of disclosed but unpatched vulnerabilities at the end of 2028 exceeds the number at the end of 2026; it fails if it has shrunk, or if no such program publishes both counts.

Forecast 13.1 (Cross-lineage oversight beats self-oversight). By the end of 2028, in controlled studies at matched capability, monitors from a different model family or training lineage catch more misaligned actions than monitors from the monitored model's own family.

Horizon 2028-12-31 · probability 55% · check: Read the controlled monitoring studies published through 2028 that compare, at matched capability, monitors from the monitored model's own family with monitors from another family or lineage on held-out misaligned behavior. The forecast holds if most of them find the cross-lineage monitors catch more; it fails if most find same-family monitors as good or better, or if none is published.

Forecast 13.2 (Drift becomes a reported metric). By the end of 2027, a system card or external evaluation reports cross-generation drift as a number: the change in answers to one fixed battery of value-laden probes between a model and its successor or a preserved predecessor.

Horizon 2027-12-31 · probability 30% · check: Read the system cards and evaluator reports published through 2027. The forecast holds if one reports such a change as a metric of its own, beyond side-by-side per-model scores; it fails otherwise.

Forecast 13.3 (Alignment automates last). Through 2030, frontier labs keep human sign-off on the alignment evaluations that gate deployment, even where they report running other AI-research work end to end without it: alignment evaluation is the last research subtask they delegate.

Horizon 2030-12-31 · probability 80% · check: Read the frontier labs' safety frameworks, system cards and research reports published through 2030. The forecast fails if any lab states that an alignment evaluation gating a deployment was run and approved by agents without human sign-off; it holds otherwise.

Forecast 14.1 (The formal frontier keeps falling). By the end of 2027, AI systems resolve at least 15% of FrontierMath: Erdős under its published protocol, up from 3% in September 2026.

Horizon 2027-12-31 · probability 55% · check: Read the best protocol score of any model, public or internal, reported under the published FrontierMath: Erdős protocol by the end of 2027; the claim holds at 15% or more and fails below it.

Forecast 14.2 (Specification becomes the binding constraint). Through 2028, at most one AI-claimed resolution of a famous open problem is accepted without a substantive dispute over whether its formal statement matches the problem, such as its hypotheses, forcing terms or boundary conditions.

Horizon 2028-12-31 · probability 35% · check: List the AI-claimed resolutions announced from October 2026 through 2028 of famous problems, meaning those with a prize, those on Wikipedia's list of unsolved problems in mathematics and Erdős problems with a prize of at least \$500, with the main objection to each. Two or more accepted by the problem's maintainers or a refereed journal with no substantive dispute about the statement falsify the claim.

Forecast 14.3 (The latency bound holds in medicine). Through 2028, the FDA approves no molecule whose target and structure were both nominated by deep-learning systems, and no drug regulatory agency accepts AI-generated in-silico efficacy evidence in place of a registration trial for a new molecule.

Horizon 2028-12-31 · probability 90% · check: Search FDA approvals and regulatory decisions through 2028; either event falsifies the claim.

Forecast 14.4 (Twins meet the optimizer's curse). By the end of 2029, AI-for-materials and AI-for-molecules programs that grow their screened candidate counts tenfold or more through a frozen twin report falling experimental confirmation rates among their top-ranked candidates, while programs that retrain the twin on every synthesis do not.

Horizon 2029-12-31 · probability 45% · check: Compare published confirmation rates of top-ranked candidates across successive screens of growing size. The forecast holds if at least one frozen-twin program reports a falling rate and no retrained-twin program reports one over a comparable growth in screen size; it fails if frozen-twin programs report stable or rising rates, or if no program publishes such rates.

Forecast 15.1 (Inertia rises in relative terms). By the end of 2027, the latest edition of *Queued Up* still puts the median time from interconnection request to commercial operation for US power projects above four years, while frontier capability keeps compounding, so physical inertia grows relative to intelligence as Proposition 15.1 predicts.

Horizon 2027-12-31 · probability 85% · check: Read the median duration for projects completed in the latest year covered by the latest edition of LBNL's *Queued Up* published by the horizon. A median of four years or less falsifies the claim.

Forecast 15.2 (The electrician premium). Through 2028, wages in US electrical trades keep growing faster than the private-sector average while data-center construction spending rises, so the scarcity at the physical link of Proposition 15.1 shows up as a premium.

Horizon 2028-12-31 · probability 60% · check: Compare growth in BLS average hourly earnings for electrical contractors and other wiring installation contractors with growth in private-sector average hourly earnings, from December 2025 to the latest month published by the horizon, and check Census construction spending on data centers over the same period. The forecast holds if electrical earnings grew faster while that spending rose; it fails otherwise.

Forecast 15.3 (Protocols beat programs). By the end of 2030, countries with open public protocols (instant payments, digital identity, open finance) show larger AI-attributable productivity gains in banking or health than comparable countries without them.

Horizon 2030-12-31 · probability 35% · check: Read the cross-country studies published through 2030 by the OECD, IMF, BIS or World Bank, or in peer-reviewed journals, that attribute productivity or cost-to-serve gains in banking or health to AI. The forecast holds if most of those that split countries by whether such protocols exist find the larger gains where they do; it fails if most find the larger gains where they do not or no difference, or if none makes the split.

Forecast 15.4 (The Baumol inversion in physical services). By the end of 2035, humanoids generalize and the relative price of labor-intensive physical services stops rising: world humanoid shipments pass one million a year, and day care, home care and cleaning no longer outpace the all-items CPI by more than a percentage point a year.

Horizon 2036-06-30 · probability 30% · check: Compare the average yearly change from 2031 to 2035 in the CPI indexes for day care and preschool, home health care and, where BLS publishes it, domestic services with that of the all-items CPI, and read world humanoid shipments from IDC or the IFR. The forecast holds if shipments passed one million in at least one year from 2031 to 2035 and those services outpaced the all-items CPI by one percentage point a year or less; it fails otherwise.

Forecast 16.1 (Verification becomes a layer). By the end of 2028, verification capacity (expert review, evaluators, test infrastructure, audit) becomes a priced, low-elasticity layer of the AI stack, and its suppliers report rising margins or order books.

Horizon 2029-03-31 · probability 50% · check: Read the annual results for 2026 and 2028 of listed suppliers of AI evaluation, expert review and data annotation, such as Innodata and Appen, published by the horizon, and the reported figures of private ones, such as Scale AI, Surge AI and Mercor. The forecast holds if at least two report higher gross margins or larger contracted backlogs for 2028 than for 2026; it fails otherwise.

Forecast 16.2 (Memory turns first). By the end of 2028, memory is the first layer of the stack whose markup cycles down as new capacity arrives: DRAM contract prices fall for two consecutive quarters before any of the big four cuts its capex guidance.

Horizon 2028-12-31 · probability 50% · check: Read TrendForce's quarterly DRAM contract prices and the capex guidance of Alphabet, Amazon, Meta and Microsoft through 2028. The forecast holds if contract prices fall for two consecutive quarters before any of the four lowers its guidance for 2027 or 2028; it fails if a guidance cut comes first, or if contract prices have not fallen for two consecutive quarters by the horizon.

Forecast 16.3 (The constraint migrates). By the end of 2029, the binding constraint of the AI stack has moved from power equipment and packaging (2026–27) to lithography and memory (2028–29), and then to the cost of capital, with rents migrating in the same order.

Horizon 2029-12-31 · probability 20% · check: The forecast holds only if the steps arrive in order: at the end of 2027, turbine and transformer lead times are still above three years or packaging still limits accelerator shipments; by the end of 2028, those lead times are below three years while DRAM or HBM contract prices are still rising or EUV tool output is the reported limit on accelerator supply; and during 2029, hyperscaler ten-year credit spreads widen while their issuance keeps growing. It fails otherwise.

Forecast 17.1 (Incidence through hiring and wages). By the end of 2030, the labor share of income in the most AI-exposed US service industries falls by several points while unemployment stays in its normal range, because the incidence runs through non-hiring and the wage cap of Proposition 16.3 more than through layoffs.

Horizon 2031-12-31 · probability 30% · check: In the BEA's GDP-by-industry accounts, compensation of employees over value added for information, finance and insurance, and professional, scientific and technical services, combined, is at least 3 points lower in 2030 than in 2025, and the annual US unemployment rate stays below 6% in every year from 2026 to 2030. A smaller fall, a stable or rising share, or any year at 6% or more falsifies it.

Forecast 17.2 (Equity before cash). By the end of 2030, the first program that pays at least a million people a dividend justified by AI has started paying, and it pays from equity stakes or fund returns, not as a cash transfer financed by taxes or deficits.

Horizon 2030-12-31 · probability 10% · check: Find the first such program by date of first payment. True if it exists and its payments come from an equity stake in AI companies, a levy paid in shares, or the returns of a public fund; false if no such program pays by the horizon, or if the first is a tax- or deficit-financed cash transfer, such as a federal "universal high income".

Forecast 17.3 (An AI dividend starts paying). By the end of 2030, a program somewhere pays at least a million people a regular dividend that its law or official announcement justifies by AI.

Horizon 2030-12-31 · probability 20% · check: Search statutes, government announcements and press coverage through 2030. True if such a program has made at least one payment to a million or more people by the horizon; false otherwise.

Forecast 17.4 (Human bottlenecks cycle). Through 2028, the scarcest human inputs of the build-out and of deployment (electricians, expert reviewers, forward-deployed engineers) keep pay growth above the economy-wide average over the period, and at least one of them passes a peak in posted pay and turns down as entrants arrive.

Horizon 2028-12-31 · probability 35% · check: Compare growth from 2025 to the latest 2028 data in BLS average hourly earnings for electrical contractors and other wiring installation contractors, and in posted pay on Indeed or Levels.fyi for forward-deployed engineers and AI-output reviewers, with growth in private-sector average hourly earnings. The forecast holds if all three grew faster than that average over the period and at least one has posted pay below its own earlier peak at the horizon; it fails otherwise.

Forecast 18.1 (The slow factors set the order). By the end of 2031, the depth of AI adoption across US occupations follows verifiability and inertia more than benchmark scores: legal and healthcare-practitioner occupations, where models pass licensing exams but outputs lack cheap verifiers and face regulation, rank below computer and mathematical occupations and below business and financial operations in the share of work hours AI assists.

Horizon 2031-12-31 · probability 50% · check: Use the occupation breakdown in the latest report of the St. Louis Fed's survey of AI use at work, or its successor, published by the horizon. The forecast holds if the share of work hours AI assists is lower for both legal occupations and healthcare practitioners than for both computer and mathematical occupations and business and financial operations; it fails otherwise.

Forecast 18.2 (The residue is where wages move). By the end of 2030, the ratio of the US median wage of electricians to that of software developers is higher than it was in 2024.

Horizon 2030-12-31 · probability 65% · check: Compare the national median annual wages of electricians and of software developers in the latest release of the BLS Occupational Employment and Wage Statistics available at the horizon with the May 2024 estimates. The forecast holds if the ratio of the electricians' median to the software developers' median is higher; it fails otherwise.

Forecast 18.3 (The residue has a price index). By the end of 2030, an index of the prices of the residue's inputs, the hard-to-verify, hard-to-trust and physical services, has risen faster than consumer prices since 2026.

Horizon 2030-12-31 · probability 65% · check: Average the BLS producer price indexes for legal services, offices of certified public accountants, electrical contractors for nonresidential building work and non-auto liability insurance, with equal weights and each rebased to its 2026 annual average, and compare the latest month published by the horizon with CPI-U rebased the same way. False if the index has risen less than CPI-U while AI deployment grew: the slow inputs would then be elastic and the residue would carry no rent.

Forecast 18.4 (Physics is more forecastable than prices). By the end of 2028, forecasts made in 2026 of the gigawatts of AI data centers for the end of 2027 and of 2028 have proved more accurate, in proportional error, than forecasts made in 2026 of AI capital spending and of the leading labs' valuations for the same dates.

Horizon 2029-06-30 · probability 50% · check: Take as forecasts the projections published in 2026 for the end of 2027 and of 2028: world AI data-center capacity from SemiAnalysis and Epoch AI; the big four's combined capital spending from bank and analyst consensus; and OpenAI's and Anthropic's valuations from bank and analyst projections, or, failing those, their last priced valuations of 2026. Compute each group's median absolute log error against the outcomes reported by the horizon. False if either the capital-spending or the valuation group has the smaller error.

Forecast 18.5 (Calibration becomes a priced asset). By the end of 2029, at least one major AI lab or regulatory authority makes a scored forecasting record or a measured rate of deference an explicit condition for deploying a model.

Horizon 2029-12-31 · probability 50% · check: Search published safety frameworks, system cards and regulatory rules. True if one makes either input an explicit condition for deployment and names the measured quantity, such as the Brier score of the lab's own capability forecasts or the rate at which a model complies with shutdown and oversight instructions in evaluations, whether or not it publishes the threshold; false if none does.

References

1. Business Trends and Outlook Survey. <https://www.census.gov/hfp/btos>
2. American Time Use Survey. <https://www.bls.gov/tus/>
3. BLS Occupational Employment and Wage Statistics. <https://www.bls.gov/oes/>