

Choosing AI That Lasts

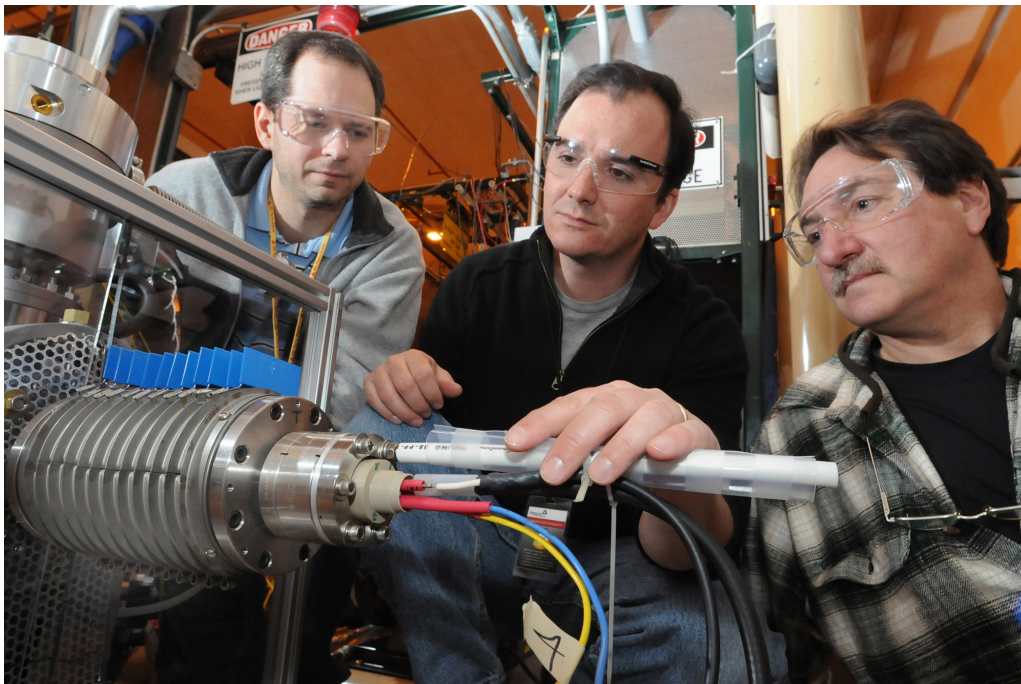
Capability, Cost, Cyber Resilience
and Survival, 2026–2031

Vikrant Salar

Founder & Principal | SALAR Technologies LLC

PUBLIC WHITE PAPER · DRAFT v2.0 · 15 PAGES

26 October 2026



Argonne scientists Bob Scott, Rick Vondrasek and Gary Zinkann inspect an ion source at the ATLAS facility, 2008. The photograph illustrates collaborative engineering and verification. Photo: Argonne National Laboratory, CC BY-SA 2.0, via Wikimedia Commons.



A practical answer to a crowded AI market

Choosing an artificial intelligence system has become an engineering decision that spans model capability, evidence quality, operating cost, security, hardware, human review and dependence on suppliers. A convincing answer is only one intermediate output. The useful endpoint is work that satisfies its stated requirements and remains dependable as models, prices, interfaces and conditions change.

This study develops an integrated framework for selecting AI configurations and evaluating their durability through 2031. It connects constrained task selection to cost per accepted task, separates model inference from execution authority, and examines cyber response through timing, coverage and recovery models. It distinguishes corporate survival from service continuity, technical relevance, economic viability and ecosystem persistence. Those outcomes can move in different directions and should not be collapsed into one ranking.

What this study adds

Decision	Research contribution
Choose	Match a versioned configuration to the job and its hard constraints
Evaluate	Count accepted work, review effort, total cost and relevant uncertainty
Protect	Separate evidence, model inference, tool permissions and action authority
Anticipate	Distinguish company survival, product continuity and ecosystem persistence

This public edition draws selected findings and figures from the completed 156-page research study. It combines primary-source synthesis, original mathematical analysis, reproducible numerical scenarios, proposed architecture and six worked cases. The numerical scenarios are explicitly assumed examples. They do not represent a new commercial-provider benchmark or fitted company-solvency estimates.

The central position is conditional: useful capability, distribution, sustainable delivery, maintained evidence and credible recovery can support endurance at several market layers. Easy substitution and persistent unpriced rework create pressure. The position can be challenged by task-level deployment results rather than accepted as a prediction of a single winner.



Start with the job you need to finish

A model name is only part of a comparison. The configuration also includes prompts, evidence, tools, permissions, hardware and the acceptance rubric. First exclude configurations that cannot meet mandatory requirements. Then compare the feasible alternatives on accepted outcomes, operating effort, latency and full cost. [1]

Job	Configuration to evaluate	Acceptance evidence
Current research	Source-enabled assistant with primary retrieval	Claims supported by relevant, dated sources
Repository repair	Coding agent in an isolated checkout	Required behavior and independent regressions
Exact numerical work	Model for framing; deterministic computation	Inputs, units and reproducible calculations
Long documents	Retrieval or long-context configuration	Correct evidence locations and critical omissions
Sensitive / offline work	Approved hosted tenant or local path	Data controls and task quality both satisfied
Routine extraction	Parser or small model first	Schema and independently labeled fields
External action	Bounded agent with separate authority	Authorized target, effect receipt and recovery

Historical coding and document benchmarks help design a relevant test set. They do not establish the best configuration for a different workload. A long context window, fluent explanation or a large test count is useful only when it connects to the acceptance criteria. [2,3,7]

A simple single-provider workflow can be the best answer. Add routing, a second model or a local fallback when a measured advantage exceeds the integration and review cost.



People and physical systems still determine delivery



The Frontier supercomputer at Oak Ridge National Laboratory, photographed in July 2022. The equipment floor shows the physical systems behind large-scale computation. Photo: Oak Ridge National Laboratory, CC BY 2.0, via [Wikimedia Commons](#).

The Frontier equipment floor at Oak Ridge illustrates the physical systems behind large-scale computation. Memory, interconnects, power, cooling and utilization affect what a model costs to serve. The photograph supplies infrastructure context; the study's energy examples use separately declared assumptions. [16]

Human review is another real operating resource. An inexpensive response may become expensive when a skilled engineer must reconstruct evidence or correct the work. The useful question is whether a configuration reduces complete task effort while preserving the required quality. Historical productivity findings remain tied to their participants, tasks and period. [25,26]

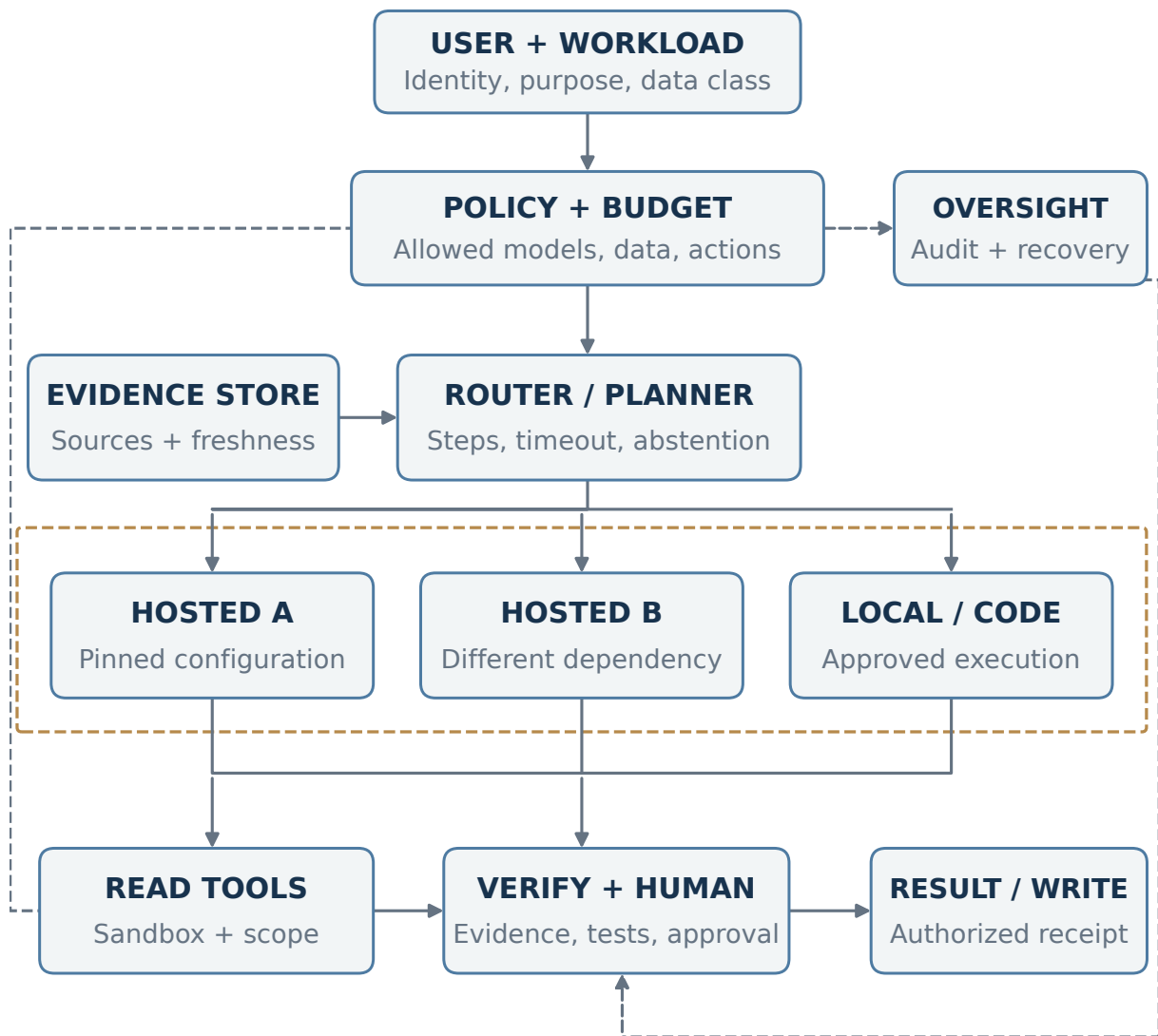
Boundary	Measure
Physical	Sustained throughput, memory, energy boundary and utilization
Human	Review minutes, rework, interruptions and accountable acceptance
Operational	Identity, evidence availability, effective recovery and support

That combination explains why neither model size nor low token price is a sufficient procurement rule. The system around the model can preserve its value or consume it.



The architecture keeps intelligence and authority clear

The proposed reference architecture begins with an authenticated task and ends with a verifiable result or an explicit refusal. It separates five functions: admission, evidence collection, inference, verification, and action execution. The distinction allows a buyer to replace a model without making the model the owner of security policy. It also allows a failed task to produce a clear receipt rather than an invented success narrative.



Gold boundary: inference. Dotted links: control / telemetry. Writes follow authorization.

Figure 1. Selected from the full study: proposed logical architecture. Evidence informs inference; independent policy determines permitted effects. Read tools and consequential writes cross different enforcement boundaries.



Trusted evidence is work a reader can inspect

Suppose the assistant reports that two regulators have equivalent current capability. A useful record contains the exact ratings, their test conditions, temperature limits, source revision, and the conversion used to compare them. A page number alone is insufficient when it points to a typical-characteristic graph instead of a guaranteed limit. The transformation from source to recommendation is part of the evidence.

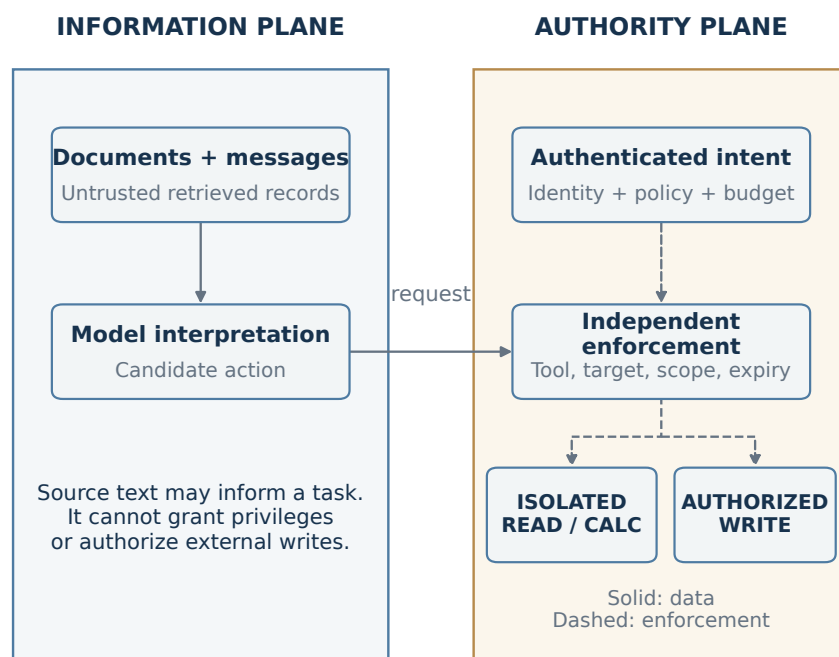


Figure 2. Selected from the full study: proposed trust boundaries. Untrusted documents and model text do not grant permissions. The data path and authority path have different responsibilities.

An answer should preserve the source revision, supporting passage, transformations and review decision for each material claim. Contradictory evidence remains visible. A blank field does not silently become zero, and several repetitions of a statement do not establish independent support. [6,8,9]

A practical system also needs source-specific access rules, bounded credentials, tool isolation and a separate check before consequential actions. NIST's AI-risk and zero-trust guidance inform those responsibilities. The proposed design still needs validation in its intended deployment. [10,15]



Count the full cost of an accepted result

Cost per accepted task = total delivery expense / independently accepted tasks

(1)

In the study's hypothetical 10,000-request example, direct routing costs \$82.70 versus \$210 for an all-large-model path. Adding \$100 fixed operations and assumed acceptance rates of 92% versus 95% reduces the useful-output advantage to 10.16%. A small-first cascade has different accounting because every request pays for its initial call. Routing research motivates testing this tradeoff, without guaranteeing a new deployment's saving. [4,5]

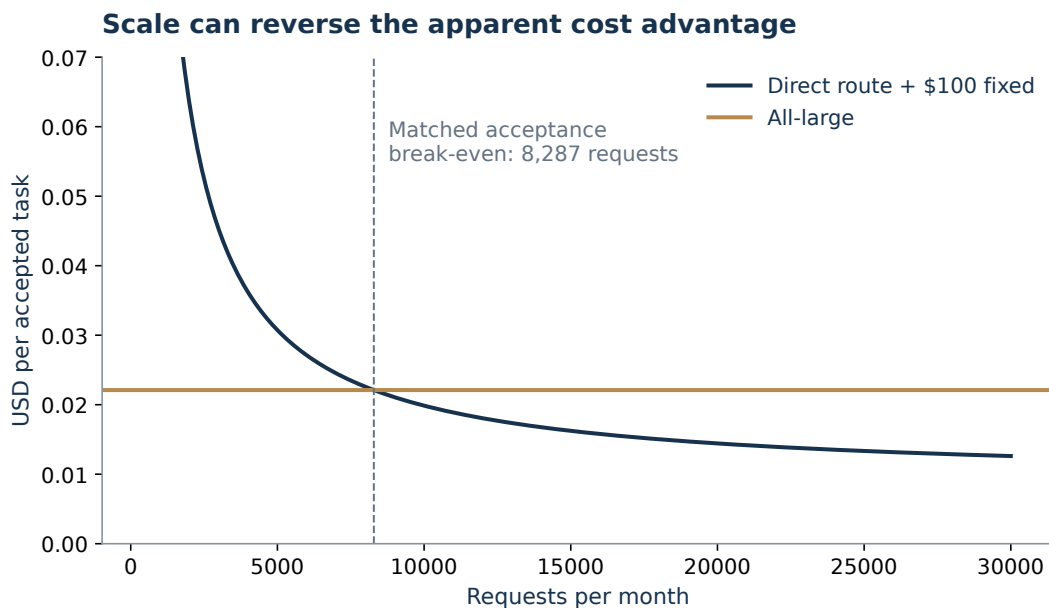


Figure 3. Selected from the full study: illustrative volume sensitivity. Prices, acceptance fractions and fixed expense are assumptions. The calculation compares total expense per accepted task.

Human review can reverse the saving entirely. In the constructed small-firm case, reducing model expense from \$21 to \$8.27 while review rises from three to four minutes increases total monthly cost from \$3,116 to \$4,103.27. Improving review to two minutes instead lowers the routed total to \$2,103.27. At \$60/hour labor, review changes dominate the model bill.



The cyber race is about effective protection

AI-assisted discovery and misuse of software weaknesses create pressure on defenders, while AI can also assist triage and repair. Vendor incident reports provide attributed observations rather than a universal rate of autonomous compromise. A repair becomes useful when it is validated and effective on the exposed asset. [13,11]

$$T_{\text{def}} = \text{detection} + \text{triage} + \text{engineering} + \text{testing} + \text{approval} + \text{effective rollout}$$

(2)

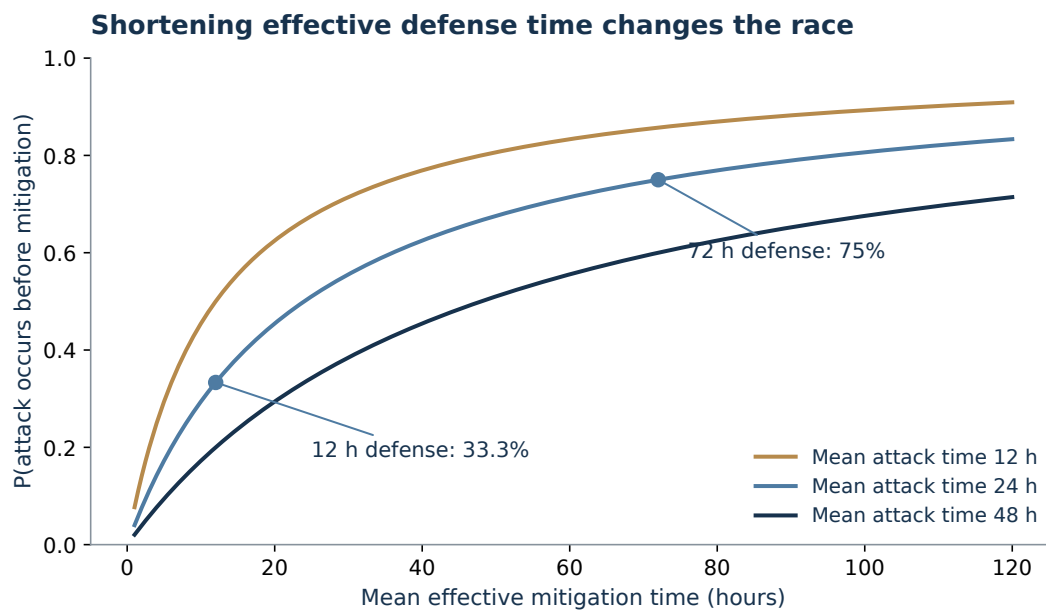


Figure 4. Selected from the full study: independent exponential timing model with declared hypothetical means. It isolates timing; the curves are not observed incident rates.

If attack and defense times are independent exponentials, the chance attack occurs first is $\lambda_a / (\lambda_a + \lambda_d)$. A mean attack time of 24 hours and defense time of 72 hours imply 75%; reducing defense mean to twelve hours gives 33.33%. The gain is conditional on the simplified race. Exposure, exploit success, patch applicability and containment are separate.

The full study extends the timing discussion to critical paths, queues and gradual rollout. FIRST's EPSS concerns exploitation of a published CVE over the next thirty days; it is not the breach probability of a particular organization. [12]



More models can share the same mistake

Two models can fail together because they share sources, training patterns, prompts, missing context, or the same flawed tool. For equal Bernoulli error probabilities e , their joint error is $e^2 + \rho e(1-e)$, where ρ is the feasible Pearson correlation between the two error indicators. This identity follows from the definition of covariance. It does not require that the models be from the same provider.

$$P(E_1 \cap E_2) = e^2 + \rho e(1-e)$$

(3)

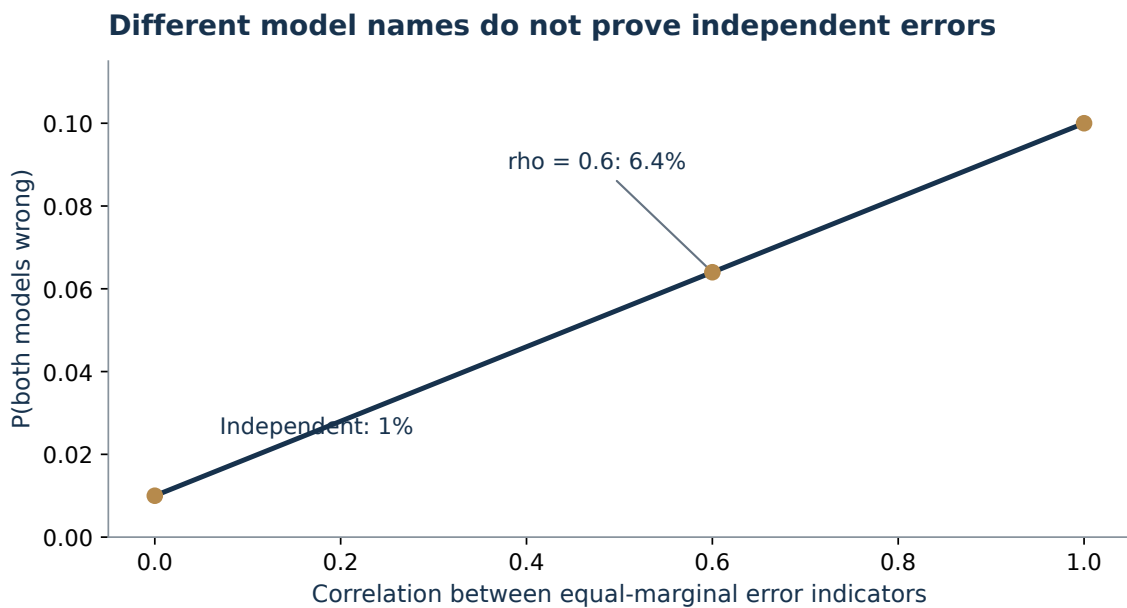


Figure 5. Selected from the full study: equal marginal error $e=0.10$. Joint error is 1% under independence and 6.4% at correlation 0.60. These are analytical illustrations, not measured model comparisons.

Joint error is not the final error rate of a voting or verification policy. The adjudicator may rescue a wrong answer, reject a correct answer or introduce a new failure. Shared retrieval, model lineage, cloud region or credentials can produce common failures even when brand names differ.

Compare strong single-model, routed, cascaded and independently verified configurations on the same held-out tasks. Fix budgets, tools and acceptance criteria. Include timeouts, contradictory evidence, abstentions and inappropriate tool actions. Measure useful coverage and reviewer effort alongside accuracy. [1,14]



Which AI business models have a path to endurance?

The study's base case is a durable, layered AI market. Platforms with distribution and dependable operating relationships can endure. Frontier services can endure when buyers continue to value their capability. Maintained open-weight ecosystems can persist through many hosting arrangements. Specialists can endure when their data, integration, or verified workflow advantage remains difficult to substitute. These are conditional pathways, not guarantees of corporate solvency.

Category and examples	Survival pathway	Pressure to monitor
Integrated platforms: Microsoft, Google, AWS	Distribution and useful operating integration	Full cost, concentration and customer value
Model services: OpenAI, Anthropic, Grok	Differentiated capability and dependable products	Repeatable task advantage and sustainable delivery
Open weights: Meta/Llama, Qwen, DeepSeek, Mistral	Adaptable artifacts and maintained hosting ecosystems	License scope, serving economics and maintenance
Operational specialists: AIP and domain vendors	Distinct data and trusted workflow integration	Implementation cost and ease of substitution

These examples identify overlapping public roles, not company failure rankings. Official documentation establishes offerings; private financial health needs additional evidence. [17,18,19,20]

The most exposed proposition is expensive delivery with no meaningful advantage that customers continue to use. A generic interface can still succeed if it earns distribution or solves a distinct workflow. A large provider can still fail a buyer's continuity needs. The study distinguishes corporate survival, service continuity and ecosystem persistence before forecasting any of them.



Four scenarios through 2031

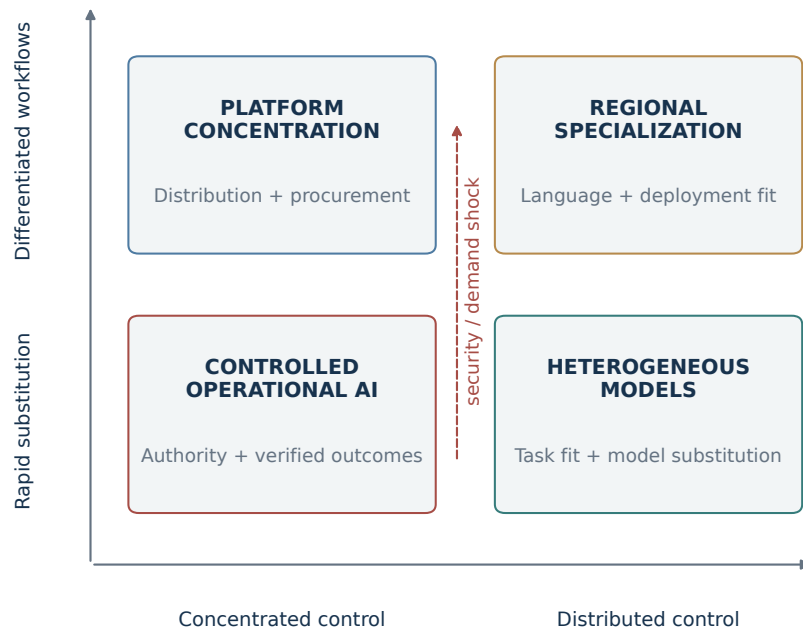


Figure 6. Selected from the full study: conceptual scenario matrix. Positions express mechanisms, not measured coordinates or probabilities. More than one scenario can develop at the same time.

Scenario	Evidence that would strengthen it
Platform concentration	Bundled procurement and integration improve accepted work at sustainable cost
Heterogeneous models	Different configurations win different tasks after all overhead is counted
Regional specialization	Local control, language and verified constraints support useful alternatives
Controlled operational AI	Higher-impact workflows reward dependable authority, evidence and recovery

Security incidents, demand shocks and financing conditions can change the pace or balance. The detailed edition sets horizons and reversal indicators, and explains competing-risk methods for distinguishing acquisition, closure and service discontinuation. It fits no private-company survival curves. [21,22,23,24]



A practical path from interest to evidence

Begin with a small task portfolio and one accountable owner. The first investment is a usable baseline: what is accepted today, how long it takes, what it costs, and which constraints are mandatory. An evaluation set and evidence record remain valuable even when the best initial choice is to keep the current workflow.

Stage	Work	Exit evidence
First 30 days	Inventory tasks, data, costs and dependencies; freeze an acceptance rubric	Versioned cases and measured baseline
Days 31-60	Compare a small number of alternatives on held-out work; include a controlled failure	Paired outcome, cost, review and recovery evidence
Days 61-90	Deploy only justified changes; rehearse export and rollback	Named owner, monitored limits and a tested exit path

The six worked cases cover a small engineering firm, controlled document review, repository repair, intermittent-connectivity edge work, defensive triage and procurement. Each translates the same principles into a different constraint set. The cases are original analytical scenarios, not undisclosed customer engagements.

A procurement decision should preserve its sensitivity. Mandatory restrictions exclude inadmissible options; weighted preferences rank feasible alternatives. When a small change in assumptions reverses the choice, say so. Portability also needs an actual exercise: export the evidence, adapt the interface and rerun the same tasks. A nominal export feature is not enough.

For a buyer, the enduring asset is a tested workflow and its evidence, rather than allegiance to one model release.



What the complete study contains



A Lawrence Livermore National Laboratory researcher holds a wafer used for research on neural-interface devices, 2013. Photo: U.S. Department of Energy, via Wikimedia Commons. Public domain.

The full 156-page edition develops the technical foundations, evidence workflow, comparative evaluation methods, cyber models, architecture, operation and recovery, full task economics, hardware and energy, provider outlook, survival methods and conditional scenarios. It includes detailed equations, original numerical outputs, twenty scientific figures and diagrams, documentary photographs, six worked cases and a 65-source bibliography.

The numerical implementation preserves assumptions and reproducible outputs. That makes the arithmetic inspectable while leaving the deployment questions open to empirical testing. A narrow simulation interval does not establish the real-world accuracy of an assumed model. New measurements can strengthen, narrow or overturn the study's conclusions.

Request the full research edition

For the full study and technical discussion, contact Vikrant Salar or SALAR Technologies LLC. The contact details follow the references. The public white paper is a draft edition prepared for research communication; it does not claim external peer review.



References

- [1] **Percy Liang et al.**. Holistic Evaluation of Language Models. 2023.
<https://arxiv.org/abs/2211.09110v2>
- [2] **Howard Yen et al.**. HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly. 2025.
<https://arxiv.org/abs/2410.02694v3>
- [3] **Carlos E. Jimenez et al.**. SWE-bench: Can Language Models Resolve Real-World GitHub Issues?. 2024.
<https://arxiv.org/abs/2310.06770v3>
- [4] **Isaac Ong et al.**. RouteLLM: Learning to Route LLMs with Preference Data. 2025.
<https://arxiv.org/abs/2406.18665v4>
- [5] **Lingjiao Chen, Matei Zaharia, and James Zou.** FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. 2023.
<https://arxiv.org/abs/2305.05176v1>
- [6] **Patrick Lewis et al.**. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020.
<https://arxiv.org/abs/2005.11401>
- [7] **Mark Chen et al.**. Evaluating Large Language Models Trained on Code. 2021.
<https://arxiv.org/abs/2107.03374v2>
- [8] **Timnit Gebru et al.**. Datasheets for Datasets. 2021.
<https://arxiv.org/abs/1803.09010v8>
- [9] **Margaret Mitchell et al.**. Model Cards for Model Reporting. 2019.
<https://arxiv.org/abs/1810.03993v2>
- [10] **C. Autio et al.**. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024.
<https://doi.org/10.6028/NIST.AI.600-1>
- [11] **M. Souppaya, K. Scarfone, and D. Dodson.** Secure Software Development Framework (SSDF) Version 1.1. NIST SP 800-218. 2022.
<https://csrc.nist.gov/pubs/sp/800/218/final>
- [12] **FIRST EPSS Special Interest Group.** Exploit Prediction Scoring System: How EPSS Works. 2026.
<https://www.first.org/epss/how-it-works>
- [13] **Anthropic.** Disrupting the first reported AI-orchestrated cyber espionage campaign. 2025.
<https://www.anthropic.com/news/disrupting-AI-espionage>



References and contact

- [14] **E. Debenedetti et al.**. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv:2406.13352v3. 2024.
<https://arxiv.org/html/2406.13352v3>
- [15] **S. Rose, O. Borchert, S. Mitchell, and S. Connelly**. Zero Trust Architecture. NIST SP 800-207. 2020.
<https://doi.org/10.6028/NIST.SP.800-207>
- [16] **International Energy Agency**. Energy and AI. 2025.
<https://www.iea.org/reports/energy-and-ai>
- [17] **Amazon Web Services**. Amazon Bedrock. 2026.
<https://aws.amazon.com/bedrock/>
- [18] **Microsoft**. What is Microsoft Foundry?. 2026.
<https://learn.microsoft.com/en-us/azure/foundry/what-is-foundry>
- [19] **Google Cloud**. Gemini platform. 2026.
<https://cloud.google.com/products/gemini-enterprise-agent-platform>
- [20] **Mistral AI**. Models Overview. 2026.
<https://docs.mistral.ai/models>
- [21] **D. R. Cox**. Regression Models and Life-Tables. 1972.
<https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- [22] **Jason P. Fine and Robert J. Gray**. A Proportional Hazards Model for the Subdistribution of a Competing Risk. 1999.
<https://doi.org/10.1080/01621459.1999.10474144>
- [23] **Glenn W. Brier**. Verification of Forecasts Expressed in Terms of Probability. 1950.
https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml
- [24] **Columbia University Mailman School of Public Health**. Competing Risk Analysis. 2026.
<https://www.publichealth.columbia.edu/research/population-health-methods/competing-risk-analysis>
- [25] **J. Becker, N. Rush, B. Barnes, and D. Rein / METR**. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. 2025.
<https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>
- [26] **METR**. We are Changing our Developer Productivity Experiment Design. 2026.
<https://metr.org/blog/2026-02-24-uplift-update/>

Full study and research discussion

Vikrant Salar | Founder & Principal
SALAR Technologies LLC
vikrant@salartechnologies.com
contact@salartechnologies.com
<https://www.salartechnologies.com/ai-in-contest.html>

