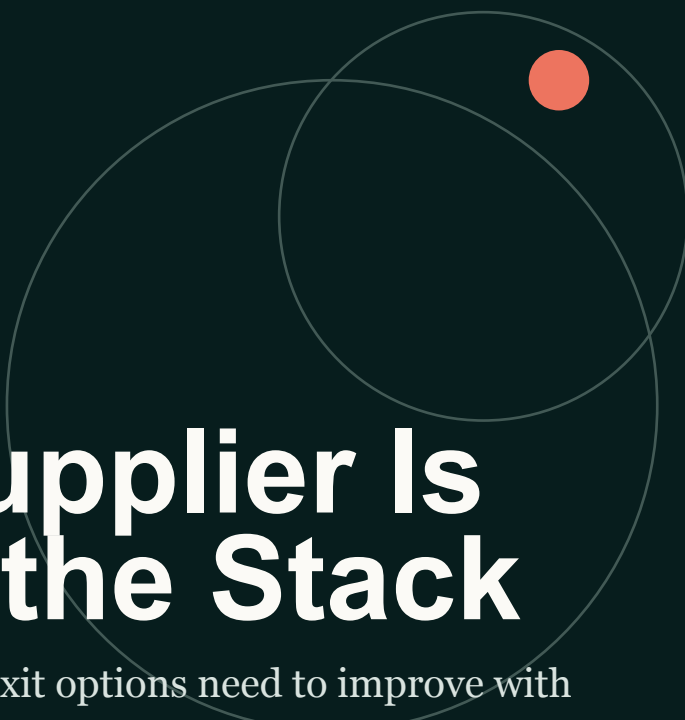


THE AI INSTITUTE / OPERATING INTELLIGENCE

CAPABILITY BRIEFING · 26 AUG / 2026



Your AI Supplier Is Becoming the Stack

Efficiency may improve quickly. Your exit options need to improve with it.

Boards, C-suite and senior management

Research period: August 2026 – August 2026

First published: 26 August 2026

Current to 26 August 2026

PUBLICATION RECORD

An Institute thesis and decision framework

As frontier AI suppliers integrate chips, models, serving software and products, buyers should evaluate AI as an operating stack and preserve portability at the data, workflow, evaluation and commercial layers.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- 01
- Measure cost per accepted business outcome, not benchmark rank or token price alone.
- 02
- Ask which benefits depend on one supplier's chip, model, serving layer and product working together.
- 03
- Keep data, workflow logic, evaluation controls and commercial exits portable enough to test.
- 04
- Run one alternate-route test before efficiency gains become structural dependency.

Research method

This capability briefing compares official supplier disclosures and deployment plans with competing cloud-platform materials, IEA infrastructure analysis, current legal and standards sources, and the newest arXiv release batch available by 26 August 2026. It distinguishes supplier benchmarks from production outcomes, initial deployments from current availability, enacted law from voluntary standards and preprints from peer-reviewed work.

First published 26 August 2026 · Updated 26 August 2026 · Research period August 2026 – August 2026 · Version 1.0 · Suggested citation: The AI Institute, *Your AI Supplier Is Becoming the Stack* (2026).

CONTENTS

Inside this paper

01	The supplier decision moved below the model The shift
02	Better inference changes which work is economical What may become possible
03	Convenience and dependency rise together The operating consequence
04	Buy the outcome, not the benchmark How to buy
05	Keep four exits practical What to preserve
06	The workflow still needs its own proof What remains unproven
07	The direction travels; the operating conditions do not What changes by region
08	Run one stack test before scaling The next 30 days
M	Method and limitations
R	References

How to use this brief

Start with the decision and use the framework to challenge the current plan. Figures retain their original populations and definitions. Institute analysis is labelled, and limitations are recorded before the references.

The supplier decision moved below the model

OpenAI has published the first measured results for Jalapeño, its custom inference chip. On three public model workloads, the company reports more work per watt, lower end-to-end latency and higher performance for highly interactive use than leading commercially available systems in the comparison. Initial deployment is expected by the end of 2026.¹²

The specifications are not the executive story. Jalapeño is designed to sit beneath OpenAI's models, serving software and products. Google and AWS are also joining custom accelerators, model platforms and serving layers.³⁴⁵ The unit of competition is widening from a model to an operating stack.

The model may win the demonstration. The stack determines the operating economics and the exit cost.

EXHIBIT / THE FOUR-LAYER STACK TEST

1 Data Can context, records and permissions move without reconstruction? Institute test	2 Workflow Does the organisation own the process logic, tools and exception paths? Institute test	3 Evaluation Can quality, safety and acceptance tests run on an alternative route? Institute test	4 Commercial Do contracts, capacity and service terms permit a practical exit? Institute test
---	---	---	---

Better inference changes which work is economical

Inference is the work a trained model performs each time it answers, classifies, plans or acts. If that work becomes faster and more energy efficient, value can appear in ordinary operating terms: shorter wait times, more attempts within a budget, and longer workflows that remain economical. OpenAI reports 1.5–1.9 times more work per watt at peak throughput, 1.7–3.6 times lower end-to-end latency and 2.1–4.1 times higher performance for highly interactive workloads in its tests.¹

Those results are supplier-authored and cover selected open models on public workloads. They do not establish production reliability, customer pricing, regional availability or a universal advantage. The right response is a bounded business test: compare the total cost and elapsed time required to produce work that an accountable owner accepts.

Convenience and dependency rise together

A tightly integrated stack can be easier to buy and operate. The supplier can tune chips, models and software as one system, remove hand-offs and expose a single commercial surface. That simplicity has value.

It also changes the failure and bargaining surface. A pricing change, capacity shortage, regional restriction, service incident or product redesign can now affect several operating layers at once. Performance achieved through the complete stack may not transfer cleanly to another supplier. Boards should therefore price dependency as part of the return, not treat it as a later architecture concern.

Buy the outcome, not the benchmark

A benchmark can reveal direction, but it rarely represents the complete business process. Ask what it costs to produce one accepted analysis, resolved case, approved design or completed customer interaction. Include model and infrastructure charges, retries, human review, integration, exception handling, controls and delay.

Then ask which part of that result depends on the supplier's integration. A low unit price can be rational even with material dependency. The discipline is to make the trade visible and compare it with a credible alternative before approving scale. The IEA's view of accelerating capital expenditure alongside rapidly falling energy per task is a reminder that efficiency can expand total use rather than reduce infrastructure demand. ⁶

Keep four exits practical

First, keep the organisation's data context, permissions and retention rules separable from the supplier's interface. Second, own the workflow definition: the business rules, tool connections, escalation paths and exception handling. Third, retain the evaluations that decide whether work is acceptable, safe and lawful. Fourth, negotiate commercial terms that make capacity, export and transition possible in practice.

Portability does not mean changing suppliers every quarter. It means being able to test an alternate route, negotiate with credible information and respond when risk or economics changes. The useful proof is a working test, not a clause that has never been exercised.

The workflow still needs its own proof

The newest research batch shows why faster infrastructure cannot replace operating assurance. In SWE Refactor Bench, only 28 of 520 agent runs completed a complex software migration while passing behavioural tests and an independent agentic audit. Another preprint finds that sharing full solutions among multiple agents can remove useful diversity. A third demonstrates that one interaction can poison an agent's persistent memory.⁷⁸⁹

These are version-one preprints, not production verdicts. Their shared executive meaning is practical: verify the whole workflow, design information boundaries deliberately and treat persistent memory as a security surface. Faster execution makes those controls more important, not less.

The direction travels; the operating conditions do not

Full-stack cloud capacity is deepest in major data-centre markets. Elsewhere, network quality, power availability, language coverage and procurement capability can dominate model performance. Data-localisation and sovereignty requirements may restrict which stack layers can be used or where they run. A global contract does not guarantee equivalent regional economics or service.

The European Union now has enforced AI Act provisions and an operational enforcement framework.¹⁰ NIST is advancing voluntary interoperability and security standards for agents.¹¹ Australia's new assurance framework applies to government and its suppliers; it is a useful public-sector comparison, not a general private-sector rule or a global proxy.¹² Each region needs its own availability, legal and exit test.

Run one stack test before scaling

Choose one important workflow that already has a clear acceptance test. Record its complete cost, delay, human review and failure path on the preferred stack. Then run the smallest credible alternative: a different model, accelerator route or serving layer using the same data boundaries and evaluations.

The purpose is not to crown a permanent winner. It is to discover what is genuinely portable, what performance depends on integration and which dependency the organisation is willing to fund. Assign an operating owner, a commercial owner and a risk owner. Bring the result back as a decision about value and resilience, not as a technical comparison.

The 30-day test — one accepted outcome, one alternate route, four proven exits and three named owners.

RESEARCH RECORD

Method and limitations

Method

This capability briefing compares official supplier disclosures and deployment plans with competing cloud-platform materials, IEA infrastructure analysis, current legal and standards sources, and the newest arXiv release batch available by 26 August 2026. It distinguishes supplier benchmarks from production outcomes, initial deployments from current availability, enacted law from voluntary standards and preprints from peer-reviewed work.

Limitations

Jalapeño has not yet entered reported production deployment, and its published results are supplier-authored. Customer pricing, regional capacity, service reliability and independent replication are not available. The cited arXiv papers are version-one preprints. Regional legal and procurement conditions require local advice. This briefing is strategic operating guidance, not technical, legal or financial advice.



Sources and limitations

Source facts link to the original publication. Institute analysis reconciles definitions and turns the material into a management proposition. Frameworks are decision aids, not claims of universal causality. Read every important statistic with its population, date, definition and caveat.

Research cutoff: 26 August 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

01 OpenAI, Jalapeño's first results

Supplier benchmark published 25 August 2026; production deployment and independent replication remain pending.

02 OpenAI and Broadcom, Jalapeño inference chip

Official announcement dated 24 June 2026; initial deployment expected by end-2026.

03 OpenAI, The full stack behind abundant intelligence

Supplier strategy statement published 25 August 2026.

04 Google Cloud, AI infrastructure at Next 26

Official platform disclosure published 23 April 2026; includes custom accelerators and open serving layers.

05 AWS and Red Hat, open-source AI infrastructure

Official supplier material; used to compare platform direction, not performance.

06 IEA, Key Questions on Energy and AI

Intergovernmental analysis published 16 April 2026; projections remain uncertain.

07 arXiv, SWE Refactor Bench

Version-one preprint submitted 24 August 2026; not peer reviewed.

08 arXiv, The Interaction Tax

Version-one preprint submitted 24 August 2026; controlled experiments.

09 arXiv, InjecMEM

Version-one preprint submitted 24 August 2026; attack setting may not generalise.

10 European Commission, Enforcement of the AI Act

Official enforcement guidance updated 24 August 2026; obligations vary by role and provision.

11 NIST, AI Agent Standards Initiative

Voluntary standards initiative; not law.

12 Australian Government, National Framework for the Assurance of AI in Government

Government assurance framework announced 24 August 2026; not a general private-sector obligation.

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Research · Readiness · Capability

theai.institute