

THE AI INSTITUTE / OPERATING INTELLIGENCE

CAPABILITY BRIEFING · 14 OCT / 2026



Decide Where AI Work Is Allowed to Run

New local-and-cloud AI tools make work placement a management decision. Set the boundary before buying the fleet.

Boards, C-suite and senior management

Research period: October 2026 – October 2026

First published: 14 October 2026

Current to 14 October 2026

PUBLICATION RECORD

An Institute thesis and decision framework

Choose where each kind of AI work may run, and what happens when that route fails, before committing to new devices or automatic cloud fallback.



WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- | | |
|----|--|
| 01 | A local model is not proof that the whole workflow stays offline. |
| 02 | Approve execution locations and fallback behaviour by task and data. |
| 03 | Test the actual route, including failures, before expanding a device fleet. |
| 04 | Compare complete service costs and local conditions, not just model speed. |

Research method

Institute synthesis of dated supplier announcements, official implementation documentation, two versioned research papers and an institutional event. Release availability is separated from planned capability; benchmark findings are separated from production outcomes. Placement categories and trial recommendations are Institute analysis.

First published 14 October 2026 · Updated 14 October 2026 · Research period October 2026 – October 2026 · Version 1.0 · Suggested citation: The AI Institute, *Decide Where AI Work Is Allowed to Run* (2026).

CONTENTS

Inside this paper

01	Buy a permitted way of working, not a label Technology & Operating Model
02	Local is not the same as offline What the product actually does
03	Write the route and the failure rule together A practical control
04	Test availability as well as capability What research adds
05	Carry the decision method, not one purchasing answer Global application
06	Run one observed placement trial The next management meeting
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.

Buy a permitted way of working, not a label

An employee asks an AI assistant to analyse a confidential file. The model may run on their laptop, another computer on the local network or a hosted service. Those choices affect who can receive the information, how the work continues during an outage and who must support it. A product described as local AI does not settle all three.

The management decision should come before a fleet purchase: which work is allowed to run where? Start with the information involved and the action the assistant can take. A public-document summary and a sensitive customer case need not have the same route. Neither should inherit its permissions simply because a new device can run the model.

On 13 October, NVIDIA announced that its Personal AI Router, PAIR, had left beta, alongside updates across its PC application ecosystem. Other elements are staged: Nemotron 3.5 Bolt is scheduled for 15 October and RTX Spark laptops for 16 October. Some integrations are expected later in the month. These dates describe supplier plans, not evidence that an enterprise deployment is ready.¹

Local is not the same as offline

PAIR can direct a request to a compatible computer on the same local network. It does not combine several computers' graphics memory or split one request across their processors. The requested model must be available on an eligible node. That distinction matters when a purchasing case assumes that several smaller devices become one larger machine.²

Local-network routing also differs from choosing between a local model and a cloud service. NVIDIA's announcement discusses both, including its Switchyard routing library. Treat them as separate capabilities with separate release dates and information boundaries, not a single promise that every application will keep every job local.¹

GitHub provides a useful, specific warning. Its 7 October Copilot CLI announcement says choosing a local model does not enable offline mode or disable telemetry. It also warns that a remote provider can still receive prompts and code context even when offline mode is enabled. Its intelligent local routing is announced with availability still forthcoming. This is evidence about that product, not every AI assistant.³

Write the route and the failure rule together

Use three simple categories for a pilot: this device, an approved local-network machine, or an approved cloud service. These are proposed management categories, not assurances supplied by a router. Record which categories each task may use, what information may travel and which person owns the decision.

Then specify what happens when the approved route cannot serve the request. A model might be missing, a machine busy or a connection unavailable. For sensitive work, the right behaviour may be to stop and explain why. For another task, an approved alternative may be acceptable. Silent fallback should not make the policy decision for you.

Follow the whole job, not only the model call. An assistant can run inference locally and still use a remote search service, repository or business application. Review those connections and the permissions to change records or send information. The useful boundary is the workflow the customer or employee actually uses.

Ask for an observable record of the destination. PAIR's documentation explicitly notes that receiving a reply is not enough to prove a request used the intended remote node. Apply that discipline more broadly: a successful answer is evidence of completion, not evidence of where processing happened. ²

Test availability as well as capability

Recent research reinforces the need to test changing conditions. RATED studies tool selection that uses both task fit and execution reliability. Its authors report improved results in a benchmark with injected API failures. That suggests a useful design direction: selection should respond to whether a service can actually perform the job. It does not establish an enterprise outage rate or a privacy guarantee. ⁴

DualQuant explores reducing the memory and computation needed for language models. Its reported improvements use particular models, tests and hardware, including an RTX 4090 implementation. Smaller local deployments may become more practical, but a research speed-up is not a complete business case. Neither study proves that your workload should move from cloud to device. ⁵

Both papers were first submitted on 9 October and were visible in the latest release batch inspected, dated 12 October. DualQuant is a preprint. RATED's authors report conference acceptance; that status should not be confused with an independent replication of their results.

Carry the decision method, not one purchasing answer

Cross-border organisations should check the actual processing destinations, supplier terms and applicable obligations for each deployment. An office in Europe, North America or Asia-Pacific may use the same assistant but a different provider configuration. Physical proximity alone does not establish permission to process data or a contractual assurance.

The economics also change by location and institution. Device supply, reliable power, connectivity, support capacity and performance in the working language belong in the trial. These are variables to measure, not reasons to assume every market in Africa, the Middle East or Latin America has the same constraint.

The World Bank's recent Small AI event frames affordable, locally adapted applications as a development opportunity. That is a useful counterweight to treating high-end hardware as the universal starting point. It is an event framework, not measured proof of job creation or local-AI returns. Comparable independent fleet economics across regions remain missing.⁶

Run one observed placement trial

Choose three real tasks with different information sensitivity and service needs. Have the business owner, IT, security and finance agree the permitted routes and failure response. Test a busy machine, a missing model and a lost connection as well as the normal path. Check destinations directly rather than relying on the answer’s quality.

Compare the complete service with the current alternative: accepted output, waiting time, human review, support effort, energy and device cost. Retain failed jobs. Expand only where the evidence supports a repeatable operating arrangement. The next purchase should follow a demonstrated way of working, not precede the decision about where work may go.

EXHIBIT / A PROPOSED PLACEMENT POLICY

DEVICE	NETWORK	CLOUD
This machine Define eligible tasks and what must remain here. Institute proposed pilot design	Approved local node Verify the destination and its availability. Institute proposed pilot design	Approved service Permit explicitly; never assume fallback is allowed. Institute proposed pilot design

RESEARCH RECORD

Method and limitations

Method

Institute synthesis of dated supplier announcements, official implementation documentation, two versioned research papers and an institutional event. Release availability is separated from planned capability; benchmark findings are separated from production outcomes. Placement categories and trial recommendations are Institute analysis.

Limitations

No independent cross-regional fleet ROI or end-to-end privacy evaluation was established. Product behaviour is configuration-specific. Research findings use bounded benchmarks; conference acceptance is author-reported for RATED. The latest visible arXiv batch remained 12 October, not a newly verified 13 October release.



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: 14 October 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

01	NVIDIA: local AI ecosystem and staged releases Gerardo Delgado, 13 October 2026. Supplier announcement; PAIR exits beta, while model, laptop and integration dates vary. Not independent enterprise ROI evidence.
02	NVIDIA: PAIR getting started Documentation inspected 14 October 2026; page publication date not stated. Local-network routing, node/model requirements, no pooled memory, and destination verification. Configuration-specific.
03	GitHub: discover local models in Copilot CLI 7 October 2026. Product-specific local discovery, telemetry and offline caveats. Intelligent routing announced with future availability; not universal behaviour of local AI.
04	Chen and Liao: RATED New submission 9 October 2026, latest inspected batch 12 October. Tool-routing benchmark with injected failures. Authors report CIKM 2026 acceptance; proceedings and independent replication not separately verified.
05	Salmani, Sharify and Mostafa: DualQuant New preprint 9 October 2026, latest inspected batch 12 October. Model- and hardware-specific quantisation evaluation; not an enterprise cost study.
06	World Bank: Small AI, Big Jobs Annual Meetings event, 11 October EDT / 12 October Bangkok, 2026. Institutional framing around local language, infrastructure and budgets; not measured job creation.

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Start with seven questions and a free readiness report. Then choose an optional 10–15 minute verbal assessment with Maya, our AI interviewer.

Start your free assessment: readiness.theai.institute

Research · Readiness · Capability

theai.institute