AI PC STRATEGY REPORT

When AI Moves Closer to the Work

The business case for on-device intelligence and flexible AI architecture

The shift

The AI PC is becoming a mainstream hardware category: a laptop or desktop with dedicated AI acceleration, usually including a neural processing unit. That definition describes capability, not value. An NPU can make sustained local inference practical, but meaningful differentiation emerges only when software applies that capacity to a role, integrates it into a real workflow and produces an outcome people can verify.

01	02	03	04	05
AI PC HARDWARE	**ON-DEVICE RUNTIME**	**ROLE-SPECIFIC SOFTWARE**	**HUMAN WORKFLOW**	**VERIFIED OUTCOME**

EVIDENCE

The NPU is enabling infrastructure. Software and workflow design determine whether it becomes productive infrastructure.

Gartner's public research distinguishes the AI PC hardware transition from the still-unproven productivity value of many current features, and argues that success requires a shift from hardware-centered positioning to software-driven, role-based solutions. [20]-[24]

What well-designed AI PC software can change

SECURITY
Raw information can be inspected, classified and transformed before any external transfer.

COST
Suitable recurring work can use already-purchased compute instead of metered inference and duplicated storage.

SPEED
No network round trip is required for local detection, search, transcription, filtering or compact-model inference.

CONTROL
The user or enterprise owns source understanding, policies, indexes, labels and release decisions.

PORTABILITY
A common preparation layer can route work across local models, cloud models and business systems.

PRODUCTIVITY
AI can assist inside the document and workflow rather than requiring a separate upload-and-prompt ritual.

Central proposition

The opportunity is not to force every model onto a laptop. It is to use on-device compute for high-frequency, data-intensive and context-sensitive workflow stages: ingestion, detection, labeling, cleaning, transformation, retrieval, lightweight reasoning, validation and orchestration. Applications can then route selected tasks to enterprise or cloud systems when those destinations add enough capability to justify the transfer, cost and control requirements.

1. Why now: the laptop hardware inflection

The AI PC is becoming a mainstream hardware category. Dedicated accelerators are no longer confined to specialist systems. Gartner defines an AI PC by the presence of dedicated AI acceleration and expects the category to become normal over time. Its more recent analysis is equally important: much of the installed hardware has yet to produce significant productivity gains, so differentiation must move toward software-defined, role-specific experiences.

EVIDENCE

AI PC adoption and AI PC value are related, but they are not the same claim.

Gartner remains positive about on-device capability and locally running small language models, while cautioning that current features often lag cloud AI in demonstrated productivity. Memory costs may also slow near-term adoption. The business case must therefore be established workflow by workflow. [20]-[24]

Representative 2026 platform signals

Platform signal	Published capability	Architectural implication	Source
Windows design point	Copilot+ PCs established a 40+ TOPS NPU baseline	Broad OEM category; OS-managed AI capabilities	[7]
AMD Ryzen AI 400	Up to 60 NPU TOPS in mobile and commercial parts	CPU + integrated GPU + NPU across laptops and mobile workstations	[4]
Intel Core Ultra Series 3	Up to 180 platform TOPS across the heterogeneous processor	CPU + GPU + NPU for consumer and commercial PCs plus specialized deployments	[5]
Qualcomm Snapdragon X2 Plus	80 NPU TOPS	High-efficiency Windows laptops with integrated AI acceleration	[6]
Apple M5 Pro / Max	Neural Engine plus GPU Neural Accelerators; up to 128 GB unified memory and 614 GB/s bandwidth	Large local models, media and professional AI workflows	[2][3]

Vendor-published specifications are not directly comparable benchmarks. TOPS is a peak arithmetic measure and does not capture model compatibility, precision, memory capacity, bandwidth, software quality, thermals or sustained performance.

The heterogeneous AI PC

CPU Control flow, parsing, rules, orchestration, general application logic.	**GPU** High-throughput generation, vision, embeddings and larger parallel workloads.	**NPU** Power-efficient sustained inference, classification, audio and background assistance.	**MEMORY + SSD** Model capacity, local indexes, working sets and fast reuse of source data.

EVIDENCE

Local AI is increasingly an operating-system capability rather than a collection of vendor-specific demos.

Windows ML can route ONNX models across NPU, GPU and CPU execution providers. Core ML similarly uses CPU, GPU and Neural Engine while optimizing memory and power. [1][3][10]

What this makes practical

- **Always-available background intelligence:** classification, transcription, document parsing and event detection can run without repeated API calls.
- **Larger private working sets:** higher memory capacity and bandwidth support local indexes, embeddings and increasingly capable compact language or multimodal models.
- **Hybrid execution:** software can decide which stage belongs on CPU, GPU, NPU, a nearby server or the cloud.
- **Offline and low-connectivity work:** core workflows continue in client sites, courtrooms, aircraft, hospitals, plants and field locations.

2. From cloud-first to placement-aware hybrid AI

Cloud AI centralized capability quickly, but it also normalized a broad architectural assumption: move the data to the model. On-device AI changes the starting question. What work can be completed where the data already resides, and what is the smallest approved representation that must move?

Two data paths

CLOUD-FIRST
Source data -> upload -> provider storage and processing -> result -> local review

Primary dependency: network, provider terms, remote capacity and export path.

PLACEMENT-AWARE HYBRID
Source data -> on-device parse/detect/prepare -> local or approved remote task -> verified result

Primary control: user or enterprise chooses every boundary crossing.

A layered architecture

Layer	Function
1. Device and identity	Secure boot, disk encryption, hardware keys, user identity, device posture
2. Local data plane	Files, operational data, local databases, sensors, mail, audio, images
3. On-device AI services	Parsing, OCR, speech, detection, classification, embeddings, compact models
4. Policy and orchestration	Permitted purposes, local/cloud routing, model selection, release gates
5. Human interaction	Review, correction, approval, exception handling and professional judgment
6. Optional destinations	Enterprise systems, collaborators, cloud AI, analytics, automation and archives
7. Evidence and telemetry	Versions, lineage, decisions, performance, cost and incident reconstruction

EVIDENCE

On-device execution does not replace the cloud. It makes data movement and workload placement explicit.

Edge-computing research provides the broader architectural foundation for latency, bandwidth, privacy and resilience benefits, alongside management, security and heterogeneity challenges. [8][9][17]

The architectural win

Once preparation and policy are separated from a particular model endpoint, an organization can use small local models for continuous work, larger local models when memory permits, a private server for shared workloads, and external frontier models for selected high-value tasks. The same source understanding and controls can serve all four.

3. Security: reduce exposure before adding controls

On-device AI can improve security by reducing the number of systems that need raw information. Role-specific software can inspect content before transfer, enforce a release policy at the point of use and keep high-value mappings or local context away from third-party services. The NPU enables efficient execution; the application supplies the detection logic, policy, review and evidence.

Security effects

Security dimension	On-device mechanism	Potential impact
Data exposure	Raw files can stay on the controlled device; approved subsets move	Fewer copies and processors with direct access
Prompt security	Sensitive fields and secrets can be detected before a prompt leaves	Prevention occurs at the user action, not only in a network log
Credential and code safety	Local scanning can identify keys, tokens, secrets and restricted code	Immediate warning or blocking without disclosing the secret to a scanning SaaS
Data residency	Processing can remain within a device, site or jurisdictional boundary	Architecture can enforce location before policy exceptions are considered
Incident containment	A disconnected or locally operating workflow can continue while external services are restricted	Reduced dependency and narrower blast radius
Continuous detection	NPUs can support sustained low-power classification or anomaly detection	Security assistance can run beside ordinary work

The controlled-release pattern

EVIDENCE

The most secure byte is often the byte that never enters an unnecessary trust domain.

On-device frameworks explicitly support local inference without a network dependency. That reduces selected transmissions, but the endpoint, model supply chain, local caches and outputs still require protection. [1][3][15]-[17]

Security responsibilities move, not disappear

- **Endpoint hardening:** local files, indexes, models and mappings need encryption, identity controls, updates and secure deletion.
- **Model provenance:** downloaded models and runtimes require signing, approved sources, version control and rollback.
- **Policy consistency:** enterprises need centrally managed rules and evidence across a distributed fleet.
- **Derived-data protection:** embeddings, prompts, logs and outputs can remain sensitive even when raw files do not leave.

4. The local data engine: detect, clean, label and prepare

Before AI can reason over enterprise or professional information, the information must be made usable. This preparation work is often more repetitive, more privacy-sensitive and more deterministic than the final reasoning task. It is therefore one of the strongest opportunities for AI PC software.

01 INGEST	02 PARSE	03 DETECT	04 LABEL	05 CLEAN	06 WRANGLE	07 INDEX	08 RELEASE

Capabilities in the data engine

Function	Work performed	Result
Ingest and parse	Documents, spreadsheets, email, images, audio and local databases	Text, tables, structure, metadata and explicit failure states
Detect	PII, PHI, payment data, secrets, confidential terms, anomalies and organization-specific fields	Candidate spans, categories and confidence
Label	Sensitivity, purpose, owner, retention, quality, matter, client, location or permitted destination	Reusable semantic and governance metadata
Clean	Duplicates, malformed values, encoding, dates, missingness, hidden content and inconsistent units	Higher-quality data without moving the full source
Wrangle	Join, reshape, filter, normalize, aggregate and select fields	Purpose-built datasets for the next task
Transform	Redact, pseudonymize, tokenize, generalize or suppress	Controlled disclosure while preserving useful structure
Index and retrieve	Chunk, embed and build local search structures	Private semantic retrieval and source-linked context
Release	Route to a local model, enterprise system, collaborator or cloud service	Approved representation with lineage

EVIDENCE

Detection is useful because a human can verify it; preparation is valuable because the verified result can be reused.

NIST's document-annotation research emphasizes machine assistance for more efficient human categorization. Recent PII research separates detection, contextual relevance and privacy-risk decisions. [12]-[14]

Why local context matters

The worker closest to the data often knows why a field exists, whether a name is public or private, which code identifies a client, which row is an exception and what the downstream user actually needs. On-device applications can place suggestions inside that working context. The person corrects the machine once; the confirmed understanding becomes available for secure analysis, compliance evidence, workflow automation and later model use.

5. Where AI PC software improves real workflows

On-device AI becomes valuable when role-specific software removes a handoff, avoids an unnecessary copy, makes assistance continuous or turns one reviewed understanding into several downstream uses. Generic hardware capability becomes differentiated only when it is connected to a person's actual decisions, tools and accountability.

Workflow	On-device activity	Process improvement
Compliance and privacy	Discover sensitive data, validate categories, record purpose and prepare access, assessment or incident evidence	Faster inventory and review; less raw-data movement
Security	Scan prompts, files, code and local events for secrets, anomalies or restricted information	Prevention and triage at the point of action
Audit and finance	Profile ledgers, reconcile fields, detect outliers, classify documents and prepare reviewed work sets	Shorter preparation cycles and stronger source linkage
Legal and HR	Classify matters, identify parties and sensitive clauses, build local search and prepare approved extracts	Faster review while preserving matter context
Healthcare	Transcribe, structure, detect identifiers, code and prepare quality or research data	Lower disclosure and quicker human validation
Field and industrial work	Analyze sensor, image or audio data where connectivity is limited	Immediate decisions, lower bandwidth and offline continuity
Knowledge work	Local search, summarization, drafting, translation and retrieval over personal or team material	Assistance within the workflow rather than a separate upload
Software and data work	Inspect repositories, generate tests, document systems, clean datasets and run local agents	Fast iteration without sending every file to a hosted service
Media and accessibility	Speech recognition, captions, translation, image analysis and enhancement	Real-time response with lower network dependence

A reusable review loop

01	02	03	04
MACHINE SUGGESTS	**PERSON CONFIRMS**	**POLICY APPLIES**	**RESULT IS REUSED**

This loop scales from an individual consultant to an enterprise. An individual owns the device and review. A small firm shares taxonomies and escalation. An SMB adds central policies and metrics. A large enterprise distributes signed models and rules across a managed fleet while retaining human validation near the business context.

EVIDENCE

AI PC software becomes productive when it shortens the distance between information, machine assistance and accountable human judgment.

The measured outcome should be less total handling time at equal or better quality, not merely faster model inference. [12][16]

Illustrative application pattern: RangerIO

RangerIO illustrates the application layer described in this report: on-device software detects and labels sensitive information, places the findings in front of the person who understands the files, and supports a controlled path to an approved AI destination. The differentiator is not that the computer contains an NPU. It is the combination of local analysis, human confirmation, reusable understanding and workflow integration. This is a vendor-published example, not independent evidence that every deployment will produce the same outcome. [25]

6. Productivity: make AI part of work, not a destination

Cloud chat established AI as a place users visit. AI PC applications can make it a capability that accompanies files, applications and decisions. That change can remove repeated preparation and enable background assistance that would be uneconomic or intrusive if every event required a network request.

Productivity mechanisms

Mechanism	What changes	Expected effect
Continuous assistance	Local classification, transcription, retrieval or monitoring runs as work changes	Less context switching and fewer manual starts
Instant iteration	Re-run local transformations and compact-model tasks without upload or queue delay	Shorter feedback loops
Context retention	Keep local indexes, preferences, project state and reviewed labels near the user	Less repeated explanation and preparation
Expert-in-the-loop	Show suggestions directly against the source and capture corrections	Faster review with domain context
Offline continuity	Core functions remain available without reliable connectivity	Productivity in client, travel, field and restricted settings
Cross-application automation	A local agent can coordinate approved desktop files and applications	Fewer export/import steps
Selective escalation	Use a larger remote model only when the local stage identifies value	Better use of premium model capacity

Measure work, not demos

TIME
Minutes per document, queue delay, iteration time, time to approved output.

QUALITY
Detection recall, correction rate, task accuracy, rework and reviewer acceptance.

FLOW
Handoffs removed, repeated preparation avoided, offline completion and escalations.

ADOPTION
Active users, workflows reused, local versus remote routing and exception patterns.

EVIDENCE

The productivity gain can be multiplicative when one locally verified understanding supports many tasks.

A reviewed index can serve search, summarization, compliance, automation and safe external reasoning. The value comes from reuse, not only from model speed. [1][3][10][12]

Human attention becomes the scarce resource

As local inference becomes cheaper, the bottleneck moves toward review design. Systems should prioritize uncertain or high-impact items, show source context, explain downstream effects and make corrections durable. The goal is not to ask people to approve every model output. It is to focus human judgment where context changes the decision.

7. Economics: shift recurring compute into owned capacity

Most organizations already purchase laptops on a refresh cycle. AI-capable hardware turns part of that fixed asset into an inference, detection and data-preparation platform. This can reduce variable cloud demand and infrastructure duplication, particularly for high-frequency, bounded workloads.

Where cost can move

Cost area	Potential reduction mechanism	Costs that remain
Metered inference	Run classification, embeddings, retrieval, transcription or compact-model tasks locally	Local energy, packaging, model testing and support
Data transfer and storage	Avoid copying full working sets into another processing repository	Endpoint storage, backup, lifecycle and secure deletion
Central preprocessing	Clean and transform data near its owner	Fleet policy, observability and exception handling
Manual handling	Pre-highlight, prioritize and reuse validated decisions	Training, quality sampling and workflow design
Premium model use	Escalate only selected tasks or minimized context	Routing logic, evaluation and dual control planes
Dedicated infrastructure	Use existing CPU/GPU/NPU and memory for appropriate workloads	Device specification, refresh and performance impact
Migration	Keep preparation, evidence and routing independent of the destination model	Adapters, open formats and portability tests

The value equation

AVOIDED	MINUS	EQUALS
API + transfer + duplicate storage + repeated preparation + queue delay	deployment + device impact + energy + support + governance + model lifecycle	measured AI PC workflow value at equivalent quality, security and availability

EVIDENCE

Local execution can remove server dependencies and per-token charges for the work performed locally.

Windows ML describes local inference as eliminating cloud costs for that execution path; Apple's Core AI describes zero server dependencies and token costs. These are architectural mechanisms, not proof of total-cost savings. [1][3]

How to build the business case

- **Baseline:** measure current API, storage, transfer, preprocessing, user time, queue delay and migration costs.
- **Segment:** separate always-local candidates, hybrid candidates and cloud-intensive tasks.
- **Pilot:** use representative hardware, documents, languages, models and network conditions.
- **Hold quality constant:** compare costs only when the output and control objectives are equivalent.
- **Scale carefully:** include model distribution, support, monitoring, recovery and fleet variance.

8. Remove data and vendor lock-in

On-device software can change the centre of control. If the user or enterprise owns the source understanding, indexes, labels, transformations and routing policy, model providers become replaceable execution options rather than custodians of the workflow.

The portable workflow control plane

Control object	Independence gained	Design requirement
Data	Original files, structured extracts and approved derivatives remain under customer control	Documented export formats; local or chosen storage
Understanding	Labels, embeddings, mappings, corrections and lineage are reusable	Open schemas; bulk export; versioned metadata
Models	Local and remote models can be evaluated behind the same task interface	ONNX or other portable formats; routing adapters; benchmarks
Policy	Purpose, sensitivity and release rules are not trapped in a provider vault	Readable policy definitions; change history
Workflow	Human decisions and business process state survive a model change	Stable APIs; auditable event records
Exit	The organization can reconstruct service without the original provider	Tested migration, deletion and recovery exercise

Destination choice

LOCAL MODEL	ORGANIZATION SERVER	ENTERPRISE PLATFORM	CLOUD FRONTIER MODEL
Private, offline and predictable recurring work.	Shared capacity under organizational control.	Central collaboration, policy and integration.	Selected tasks needing maximum model capability.

EVIDENCE

Portability is an architecture property that must be designed and tested.

NIST's cloud standards roadmap treats data and workload portability as explicit concerns. Cross-platform runtimes such as ONNX can reduce model-format friction, but full independence also requires portable policies, metadata, mappings and review history. [10][11]

Lock-in is reduced, not abolished

Hardware accelerators, operating systems and runtimes still differ. A model that runs efficiently on one NPU may require conversion or fallback elsewhere. Local applications can create their own proprietary stores. The answer is a portability contract: named export formats, reconstructable evidence, model evaluation across targets and a tested destination change.

9. The next wave of AI PC applications

The hardware and runtime foundations support more than chat. The likely progression is from discrete local features to a persistent, policy-aware intelligence layer that understands work across applications while keeping sensitive context under user or enterprise control.

Capability horizon

Horizon	Capability	Example
NOW	Detection and classification	PII and secrets, document type, spam, anomaly and quality detection
NOW	Local preparation	OCR, transcription, translation, cleaning, wrangling, indexing and retrieval
NOW / NEXT	Private copilots	Source-linked search, drafting, summarization and task assistance over local material
NEXT	Policy-aware AI gateway	Automatic local/cloud routing based on purpose, sensitivity, cost and model capability
NEXT	Cross-application agents	Local agents coordinate files and desktop apps under explicit permissions and approval
NEXT	Adaptive personal models	On-device tuning and preference learning without centralizing raw personal work history
NEXT	Distributed enterprise intelligence	Managed models and policies execute across endpoints, branches and organization-controlled servers
EMERGING	Collaborative learning	Organizations improve models from distributed signals without collecting every raw record
EMERGING	Real-time security layer	Continuous endpoint content and behaviour analysis with immediate local response
EMERGING	Physical and field AI	Vision, audio, sensor reasoning and robotics near the event with intermittent connectivity

The converging platform

01	02	03	04	05
PERSONAL CONTEXT	**LOCAL MEMORY**	**MULTIMODAL MODELS**	**POLICY ROUTER**	**ENTERPRISE ACTION**

EVIDENCE

On-device learning extends the AI PC from inference to adaptation.

Core ML supports on-device fine-tuning, and ONNX Runtime documents local training for personalization and federated-learning scenarios. Privacy still depends on what updates, gradients or derived data leave the device. [3][18]

What will determine adoption

- More memory and sustained accelerator performance, not TOPS alone.
- Portable runtimes and model formats across heterogeneous hardware.
- Clear permissions and observable actions for local agents.
- Enterprise deployment, update, evidence and recovery controls.
- Interfaces that combine automatic assistance with focused human judgment.
- Demonstrated total-cost and workflow improvement on real work.

10. A practical adoption roadmap

Organizations do not need to replace their cloud strategy. They need to identify which workflow stages are unnecessarily remote, repetitive or disclosure-heavy, then move those stages toward the user in controlled increments.

First 90 days

Step	Action	Output
1. Inventory	Identify recurring AI and data workflows, source locations, sensitivity, current providers and variable costs	Candidate map
2. Decompose	Separate parsing, detection, preparation, retrieval, inference, review and action	Stage-level architecture
3. Select	Choose one high-frequency bounded workflow with measurable friction	Pilot charter
4. Benchmark	Test representative devices, models, formats, languages and failure cases	Quality and performance baseline
5. Design controls	Define model sources, endpoint requirements, release gates, logging, recovery and export	Control specification
6. Pilot	Run side-by-side with the current process and capture user correction	Observed productivity, cost and risk
7. Decide	Keep local, use hybrid, stay cloud-based or stop based on thresholds	Evidence-backed disposition

Workload placement guide

Disposition	Typical conditions
Prefer on-device	Sensitive raw data; frequent bounded task; low-latency or offline need; local hardware adequate; output reviewable
Prefer hybrid	Local preparation reduces context; remote model materially improves quality; transmission is approved
Prefer centralized	High concurrency; shared state; very large models; mature centralized controls; acceptable data terms
Do not automate	Purpose unclear; quality unvalidated; action too consequential; no accountable reviewer or recovery

Success measures

ARCHITECTURE
Bytes and data classes transmitted; providers with raw access; offline completion.

WORK
Cycle time, handoffs, review corrections, reuse and approved outputs.

ECONOMICS
Variable inference, storage, support and fully loaded cost per task.

CONTROL
Portability test, model provenance, policy coverage and incident reconstruction.

EVIDENCE

The decision unit is the workflow stage, not the whole application.

Moving the correct stages onto the device can produce security and economic benefits while preserving access to centralized capability where it matters. [8]-[11][16]

11. Conclusion

The AI PC transition is not simply a new generation of faster laptops. It is a redistribution of AI capability. Detection, data preparation, retrieval, compact-model reasoning and workflow orchestration can increasingly occur where people work and where information originates.

That shift has architectural consequences. Raw data can enter fewer trust domains. Security checks can occur before disclosure. Compliance and classification work can happen inside ordinary workflows. Repeated cloud calls and duplicated storage can decline. The user or enterprise can retain the data understanding and change model providers without rebuilding the entire process.

The cloud remains essential for frontier capability, centralized collaboration, high concurrency and large shared systems. The practical architecture is therefore placement-aware and hybrid: on-device for continuous, sensitive or preparatory stages where it is effective; remote by deliberate choice when a task justifies the transfer, cost and control requirements.

EVIDENCE

The AI PC becomes meaningful when software turns local compute into a role-specific, integrated and verifiable workflow.

Organizations should evaluate applications by workflow outcomes rather than accelerator specifications alone: exposure reduced, handling time removed, quality preserved, human decisions supported and destinations kept replaceable. [1]-[25]

References

Official platform sources, Gartner public material, standards and primary technical literature. Accessed August 2026.

[1] Microsoft (2026). Windows ML overview.
https://learn.microsoft.com/en-us/windows/ai/new-windows-ml/overview

[2] Apple (2026). MacBook Pro with M5 Pro and M5 Max.
https://www.apple.com/newsroom/2026/03/apple-introduces-macbook-pro-with-all-new-m5-pro-and-m5-max/

[3] Apple Developer. Core AI and Core ML for on-device models.
https://developer.apple.com/core-ai/

[4] AMD (2026). Ryzen AI 400 and PRO 400 Series.
https://www.amd.com/en/newsroom/press-releases/2026-1-5-amd-expands-ai-leadership-across-client-graphics-.html

[5] Intel (2025). Core Ultra Series 3 architecture.
https://newsroom.intel.com/client-computing/intel-unveils-panther-lake-architecture-first-ai-pc-platform-built-on-18a

[6] Qualcomm (2026). Snapdragon X2 Plus.
https://www.qualcomm.com/laptops/products/snapdragon-x2-plus

[7] Microsoft (2024). Introducing Copilot+ PCs.
https://blogs.microsoft.com/blog/2024/05/20/introducing-copilot-pcs/

[8] Shi, W. et al. (2016). Edge Computing: Vision and Challenges.
https://doi.org/10.1109/JIOT.2016.2579198

[9] Edge-AI systematic review (2025). Architectures, applications and challenges.
https://www.sciencedirect.com/science/article/abs/pii/S1084804525002723

[10] ONNX Runtime. Cross-platform inference on cloud, edge, web and mobile.
https://onnxruntime.ai/inference

[11] NIST SP 500-291r2. Cloud Computing Standards Roadmap.
https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-291r2.pdf

[12] Fung, J. et al. / NIST TN 2287 (2024). Human-in-the-loop Technical Document Annotation.
https://www.nist.gov/publications/human-loop-technical-document-annotation-developing-and-validating-system-provide

[13] Ponomarenko, M. et al. (2026). CAPID: Context-Aware PII Detection for Question-Answering Systems.
https://aclanthology.org/2026.eacl-srw.23/

[14] Papadopoulou, A. et al. (2026). Neural text sanitization with privacy risk indicators.
https://link.springer.com/article/10.1007/s10579-026-09911-1

[15] NIST. Privacy Framework.
https://www.nist.gov/privacy-framework

[16] NIST. AI Risk Management Framework 1.0.
https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

[17] ENISA (2022). Fog and Edge Computing in 5G.
https://www.enisa.europa.eu/publications/fog-and-edge-computing-in-5g

[18] ONNX Runtime. On-device training.
https://onnxruntime.ai/docs/get-started/training-on-device.html

[19] Garfinkel, S. / NISTIR 8053. De-Identification of Personal Information.
https://www.nist.gov/publications/de-identification-personal-information

[20] Gartner (2024). Gartner defines the AI PC category.
https://www.gartner.com/en/newsroom/press-releases/2024-02-07-gartner-predicts-worldwide-shipments-of-ai-pcs-and-genai-smartphones-to-total-295-million-units-in-2024

[21] Gartner (2025). AI PCs and the shift toward locally running small language models.
https://www.gartner.com/en/newsroom/press-releases/2025-08-28-gartner-says-artificial-intelligence-pcs-will-represent-31-percent-of-worldwide-pc-market-by-the-end-of-2025

[22] Gartner (2026). AI PC features have not yet delivered broad productivity gains.
https://www.gartner.com/en/newsroom/press-releases/2026-1-20-gartner-says-worldwide-pc-shipments-increased-9-point-3-percent-in-fourth-quarter-of-2025-and-9-point-1-percent-for-the-full-year

[23] Gartner (2026). Memory costs expected to slow AI PC adoption through 2027.
https://www.gartner.com/en/newsroom/press-releases/2026-02-26-gartner-says-surging-memory-costs-will-reduce-global-pc-and-smartphone-shipments-in-2026

[24] Gartner Research (2025). AI PC Success Demands a Shift From Hardware to Software.
https://www.gartner.com/en/documents/6820134

[25] RangerIO. Public user guide.
https://rangerio.com/docs

EVIDENCE

Scope statement

Hardware specifications and performance comparisons cited from vendors reflect their published configurations and methodologies. Actual model performance, energy, security, cost and workflow outcomes vary. This report is architecture analysis, not a product endorsement or legal opinion.