



ENTERPRISE AI WITHOUT COMPROMISE:

ADVANCED CAPABILITIES, COMPLETE SOVEREIGNTY, OPEN INFRASTRUCTURE

EXECUTIVE SUMMARY

Every enterprise is investing in AI. Few have moved from pilots to production at scale. Cohere and AMD offer a path onward. The North by Cohere platform, powered by AMD Instinct™ GPUs, delivers enterprise-grade agentic AI that can run entirely on-premises with zero data egress. Together, they address three requirements that enterprises need to get right: advanced AI capabilities grounded in their own data, sovereignty and regulatory compliance, and open infrastructure that scales without vendor lock-in.

THE ENTERPRISE CHALLENGE

Organizations across every industry are investing in AI pilots, proofs of concept, and innovation labs, but many struggle to move AI into production reliably at scale. The gap between experimentation and deployment is where competitive advantage is won or lost.

Common friction points include:

- Cloud API dependence can introduce data sovereignty risks, unpredictable costs, and vendor lock-in to a single provider's roadmap.
- Generic foundation models often struggle to integrate with enterprise systems, producing outputs that lack grounding, citations, and auditability.
- Regulatory mandates (GDPR, HIPAA, SOC 2, FedRAMP) can block production deployment for any solution that requires data to leave the customer's infrastructure
- Multilingual and multimodal enterprise data often remains inaccessible to AI systems built for English-only, text-only use cases.
- Scaling from a single proof-of-concept to enterprise-wide deployment can expose latency, cost, and integration problems invisible at pilot scale.

The enterprises that industrialize AI on sovereign, open infrastructure will set the pace for their industries. The question is no longer whether to deploy AI, but whether the organization has the governance, data control, and infrastructure flexibility required to move from experiments to production at scale.

A DIFFERENT APPROACH

What if you could deploy enterprise AI that is sovereign by design, grounded in your own data with verifiable citations, and built on open, cost-efficient infrastructure you control?

Cohere and AMD deliver this combination. The joint solution is built on three pillars: **sovereignty and data control** through Cohere's security-first platform with zero data egress; **grounded, auditable AI** through retrieval-augmented generation with inline citations and enterprise governance; and **open infrastructure economics** through high performance AMD Instinct™ GPUs and commitment to open standards. The sections that follow detail how each pillar works and what they deliver when combined.

THE SOLUTION: COHERE + AMD

The Cohere Application Layer

North Platform: Agentic AI Workspace

The North platform is a turnkey agentic AI workspace built for enterprise production environments. It provides pre-built agents for common enterprise scenarios (customer support, document processing, knowledge assistance, developer assistance, and operations) alongside a no-code/low-code custom agent builder for domain-specific workflows.

North integrates with the enterprise stack through plug-and-play connectors to Salesforce, ServiceNow, SharePoint, Google Workspace, Slack, Jira, Confluence, and major databases (Oracle, SQL Server, PostgreSQL, MongoDB). Custom connectors can be built using standardized APIs. Enterprise controls include guardrails, audit trails, human-in-the-loop approvals, version control, and SSO/RBAC integration.

Compass: Enterprise Search and Discovery

Built into North, Compass delivers intelligent search across scattered enterprise data. It is multimodal and multilingual, supporting 23 languages and parsing images, slides, spreadsheets, and diverse file types. Early deployments show Compass helping users complete search and discovery with 80%+ faster task completion compared to traditional search. Built on Cohere's Embed and Rerank models, Compass can be deployed in VPC or on-premises with document-level security and role-based access controls.

Command A: The Intelligence Layer

North's agents are powered by Command A, Cohere's high-performance foundation model optimized for enterprise use. Command A delivers retrieval-augmented generation (RAG) with inline citations for traceability, multi-step tool use for agentic AI workflows, and a 128K token context window for processing entire documents. Multilingual support spans 10+ languages for generation (English, French, Spanish, German, Portuguese, Italian, Japanese, Korean, Chinese, Arabic) and 23 languages for embeddings via the Aya Expanse model family. Additional model capabilities include Embed for multimodal search and retrieval, and Rerank for semantic search quality.

The AMD Infrastructure Foundation

AMD provides the compute foundation that helps enable sovereign, on-premises enterprise AI at production scale. The relationship between AMD and Cohere extends beyond a standard technology partnership: AMD Ventures is an investor in Cohere, and the companies have collaborated to optimize Cohere's enterprise models specifically for AMD Instinct™ GPUs.

AMD Instinct™ GPUs

AMD Instinct GPUs are purpose-built for the memory-intensive demands of enterprise AI. Large language models like Cohere's Command A require substantial high-bandwidth memory (HBM) capacity for long-context reasoning, retrieval-augmented generation, and multi-step agentic workflows. Instinct accelerators deliver the memory bandwidth and capacity these workloads demand, enabling enterprises to run complex models with a smaller infrastructure footprint.

Cohere's Command models are already deployed and optimized on AMD Instinct infrastructure across enterprise customers. At the AMD Advancing AI 2025 event, Cohere demonstrated its models running on Instinct accelerators for enterprise-grade LLM inference with high throughput, efficiency, and data privacy.

Cloud and On-Premises Availability

Cohere on AMD Instinct is available today through Oracle Cloud Infrastructure (OCI), with Dell server configurations in development. This gives enterprises multiple deployment paths: on-premises using OEM server platforms, through Oracle Cloud infrastructure, or in hybrid configurations. Oracle offers AMD Instinct-powered bare metal instances with enterprise-grade cluster networking, providing cloud-scale AI capacity without the overhead of virtualization. For organizations pursuing sovereign deployments, on-premises configurations deliver the same Instinct acceleration inside the customer's own data center perimeter.

Open Ecosystem and Standards

The AMD open ecosystem approach is central to the value proposition for enterprises wary of infrastructure lock-in. Instinct accelerators are supported by a broad compatibility layer spanning standard AI and MLOps toolchains including PyTorch, vLLM, and Kubernetes-native orchestration. AMD had a founding role in the UALink interconnect standard, alongside other industry leaders. This demonstrates a long-term commitment to open, multi-vendor AI infrastructure. For enterprises, this means the flexibility to evolve AI strategies, swap components, and integrate new capabilities without being locked to a single vendor's proprietary stack.

WHY THIS COMBINATION MATTERS

Cohere and AMD each bring differentiated capabilities. The value lies in what those capabilities unlock together.

What Cohere brings:

Enterprise-grade foundation models and end-to-end AI products, purpose-built for business. Co-founded by one of the authors of the seminal paper "Attention Is All You Need," which underpins today's large language models.

- Security-first architecture: private deployment options (VPC, on-prem, or fully air-gapped), data sovereignty, and audit-ready traceability.
- Production-ready agentic platform (North by Cohere) featuring Compass integration for search, embed, and re-rank capabilities, with enterprise connectors and customizable agents
- Deep multilingual capabilities: 23 languages for search, 10+ for generation.
- Strategic partnerships with SAP, Oracle, Microsoft, AMD, Fujitsu, Nvidia, Cisco, and Salesforce Ventures, demonstrating enterprise-grade AI capabilities.

What AMD brings:

- A proven compute platform for large-model inference, already running Cohere's Command models in production across enterprise customers.
- Multiple deployment paths: on-premises via OEM partners (Dell), cloud via Oracle Cloud Infrastructure, or hybrid configurations.
- An open ecosystem philosophy backed by founding participation in UALink and broad toolchain compatibility, helping enterprises avoid infrastructure lock-in.

What they create together:

Sovereign AI for regulated industries. Cohere's private deployment combined with AMD Instinct GPU on-premises infrastructure enables AI that does not need to leave the customer's perimeter. Critical for financial services, healthcare, government, and critical infrastructure.

Cost-efficient scale. The AMD open infrastructure can help lower the cost of running Cohere's models and enable more predictable economics as workloads scale.

Production-grade deployment speed. Cohere's models are optimized for AMD Instinct GPUs, reducing the integration work that can delay enterprise AI deployments. Cohere's turnkey North platform helps accelerate time from pilot to production.

Validated integration. AMD has adopted the North platform for its own internal enterprise AI workloads, making AMD both a technology partner and a reference customer.



BUSINESS OUTCOMES: INDUSTRY SPOTLIGHTS*



FINANCIAL SERVICES

THE PROBLEM

Financial institutions face mounting compliance costs and regulatory complexity across multiple jurisdictions. Manual review of regulatory documents is slow, expensive, and error-prone. Data sovereignty requirements prevent use of cloud-based AI for sensitive financial data.

THE SOLUTION

Cohere-powered compliance agents on AMD Instinct GPUs review regulatory documents with audit trails and cited sources. North's enterprise connectors integrate with existing compliance systems. Private deployment supports zero data egress. Multilingual capabilities support cross-border operations.

THE RESULT

Early deployments indicate faster regulatory review cycles and measurably reduced compliance workload, with consistent audit-ready traceability across jurisdictions.



HEALTHCARE AND LIFE SCIENCES

THE PROBLEM

Clinical and research teams need rapid access to medical databases, journals, and protocols, but data is scattered across siloed systems. HIPAA compliance requirements prevent use of external AI services. Multilingual patient populations demand language support beyond English.

THE SOLUTION

Compass indexes medical knowledge sources with multimodal, multilingual search. North by Cohere orchestrates multi-agent care coordination workflows with HIPAA-compliant on-premises deployment. Command A delivers RAG with inline citations for clinical traceability.

THE RESULT

Healthcare organizations report faster access to clinical knowledge, improved workflow coordination, and stronger patient outcomes, with full HIPAA compliance maintained throughout.



TECHNOLOGY

THE PROBLEM

Technology companies face growing pressure to improve developer productivity and reduce engineering bottlenecks. Engineering teams spend significant time navigating code repositories, technical documentation, and internal knowledge systems. Tier-1 and Tier-2 technical support remains largely manual, slowing response times and diverting senior engineering resources.

THE SOLUTION

Cohere-powered developer assistants on AMD Instinct infrastructure index code repositories, Jira, Confluence, and technical documentation for natural-language search and retrieval. North's operations agents automate tier-1/2 technical support with code troubleshooting capabilities. Multilingual support enables globally distributed engineering teams to collaborate effectively.

THE RESULT

Technology customers such as CoreWeave, Notion, Atomicwork, and DraftWise report measurable improvements in developer onboarding speed, engineering productivity, and support ticket resolution times.

DEPLOYMENT MODEL

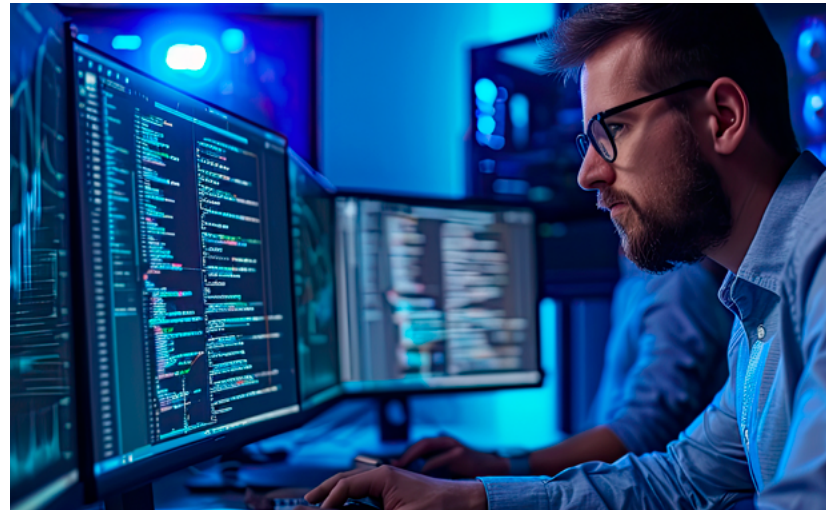
The Cohere + AMD solution supports flexible deployment models aligned with enterprise IT strategies and regulatory requirements.

Deployment Options:

- **On-premises (VPC or air-gapped):** full sovereignty, zero data egress. Cohere models deployed on AMD Instinct GPUs inside the customer's environment. Available through OEM server platforms from Dell and other partners.
- **Cloud:** Cohere models available on AMD Instinct GPU-powered instances through Oracle Cloud Infrastructure, with bare metal options for workloads requiring maximum performance without virtualization overhead.
- **Hybrid:** combine on-premises sensitive workloads with cloud-based non-sensitive workloads, maintaining sovereignty where it matters while leveraging cloud elasticity for less sensitive tasks.

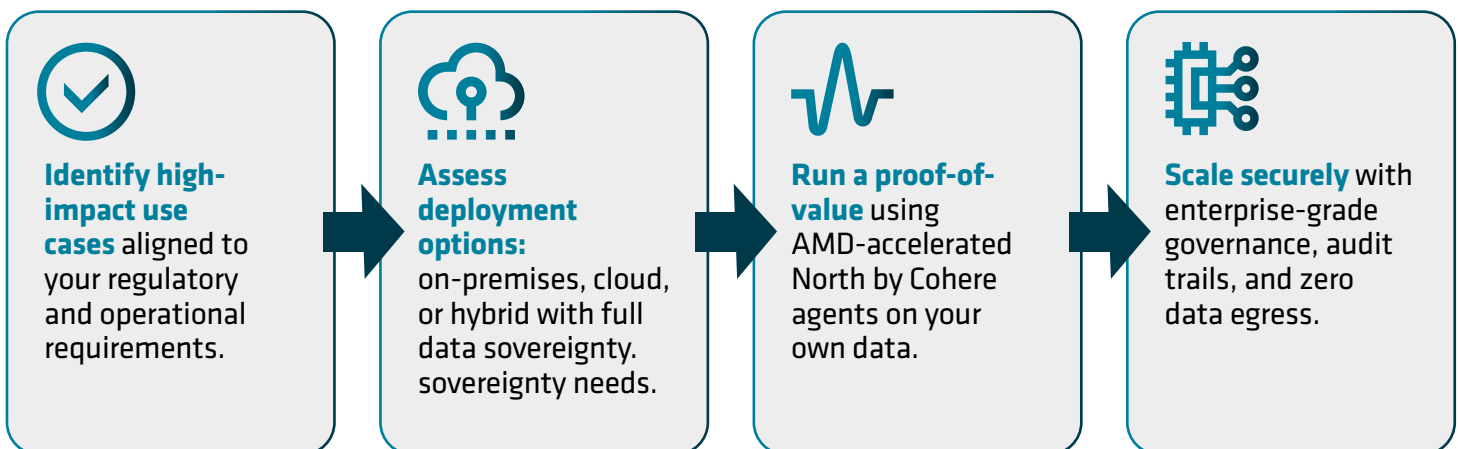
Deployment Timeline:

- **Weeks 1-2:** Pilot setup. Deploy sandbox environment.
- **Weeks 3-4:** Agent build. Develop first AI agent using North by Cohere.
- **Weeks 5-24:** Production. Full deployment with Cohere + AMD Instinct.
- **Ongoing:** Scale agents across the enterprise.



TAKE ACTION

Ready to move from AI experimentation to sovereign, production-grade AI? Engage with Cohere and AMD to:





CONTACT YOUR AMD OR COHERE REPRESENTATIVE FOR A STRATEGY SESSION.

ABOUT AMD

More than 50 years driving innovation in high-performance computing, AMD Instinct accelerators power AI workloads for leading cloud providers, enterprises, and sovereign AI initiatives worldwide. The AMD commitment to open standards and open ecosystem ensures enterprises can deploy AI with control, flexibility, and long-term vendor independence.

ABOUT COHERE

Cohere is the leading security-first enterprise AI company that builds cutting-edge foundation models and end-to-end AI products to solve real business problems. We partner with companies for seamless integration, customization, and user-friendly solutions, offering our all-in-one platform with top-tier security, data privacy, and deployment flexibility across cloud, private cloud, or on-premises environments.

*All performance and/or cost savings claims are provided by the 3rd party organization featured herein and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to such 3rd party organization and may not be typical. GD-181a.

©2026 Advanced Micro Devices, Inc. all rights reserved. AMD, the AMD arrow, and combinations thereof, are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

PID: 264651300