Mistral AI NOW Summit

Company Data (as disclosed at the summit)

Headcount	~1,000 employees (from ~15 in 2023)
2026 revenue target	€1bn (~$1.17bn); ~$400M ARR reached Jan 2026 (~20× YoY)
Customers	1,000+ high-value accounts; ~60% of revenue in Europe
Funding	≥$3.9bn across nine rounds; €1.7bn Series C led by ASML (Sept 2025, €11.7bn valuation)
Compute capital	$830M debt (Mar 2026, seven banks) for 13,800 Nvidia chips; exploring custom chips

---

1. Executive synthesis: themes across the summit

Every thread traces back to one argument the summit made explicit: **Control**. Open-weight models, on-premise or air-gapped deployment, specialised models fine-tuned on proprietary "industrial truth," and ownership of the stack from chips and compute (Koyeb, Les Ulis, possible custom silicon) up through platforms (Studio, Forge), a unified agent (Vibe), and named vertical lighthouses (Airbus, BMW, ASML; BNP Paribas, HSBC; Luxembourg, EDF).

- **Owning the full stack.** Mistral framed its strategy around controlling everything from compute and chips to applications, plus deep vertical integration for specific business use cases. There are still a lot of points to be clarified (Datacenter operations in partnership with US companies, Silicon Timeline) but this remains their stated ultimate goal.
- **Sovereignty as a design requirement.** Repeatedly defined as local hosting + local control + local expertise (Luxembourg, NTT DATA), with on-premise / air-gapped deployment recurring for regulated and sensitive data (finance, defense, public sector, automotive engineering).
- **Specialized models on proprietary "industrial truth" beat general-purpose AI.** BMW (crash testing), Ericsson (proprietary silicon), Moeve (P&ID diagrams), and the EPO (patents) all built custom models on sensitive in-house data, typically via Mistral's fine-tuning toolkit Forge.
- **The shift to agentic AI.** Moving from prompt-response chatbots to multi-step agents that plan, call tools, retrieve, reason, and act - operationalized through harnesses/scaffolds (tooling, memory, orchestration, guardrails, observability via "thinking traces").
- **Distributed / hybrid compute.** Intelligence is migrating from centralized data centers to a device–edge–cloud continuum, with orchestration deciding placement based on power, memory, latency, and data sensitivity (Qualcomm). *Power and memory bandwidth are the binding constraints.*
- **Infrastructure is now critical infrastructure.** "AI factories" optimize power, networking, compute, and data as one production system; the workload is shifting from static training to dynamic inference loops feeding fine-tuning and RL (NVIDIA / VAST / Mistral panel).
- **Mistral direct involvement with customer vs Partner delivery.** A lot of the presented cases/projects were delivered thanks to Mistral direct involvement through the deployment of "forward engineers" (lot on new hires in that area). Having spoken with partners present

at the conference (TCS, Reply, NTT DATA), everyone has stated that they are **starting** to work in the Mistral Stack, not in full delivery. Reply has formed a specific company (https://www.reply.com/sail-reply/en) branded under the “Sovereign Ai” flag with the only goal to start delivering projects by the end of the year on Mistral Stack.

2. Announcements, products & partnerships roundup

Products

- **Vibe family -** “Le Chat” is transitioning into Vibe. Vibe for Work (long-horizon agent integrating calendar/email for scheduling and summarization) and Vibe for Code (coding assistant via CLI, web, and a new VS Code extension).
- **Studio -** enterprise platform to build, test, govern, and manage AI agents/apps; supports on-premise deployment.
- **Forge -** on-site “last-mile” model-adaptation tool for fine-tuning on proprietary data, new modalities, and smaller task-specific models.
- **Models -** all now natively multimodal (folding in former Pixtral/Magistral); Mistral OCR (thousands of pages/min on one GPU); agentic search in Mistral Medium 3.5; Mistral Large 4 expected summer 2026 (fluid dynamics, cyber defense); next-gen models trained on 200+ languages; Devstral (code), MM3.5 (agentic), Voxtral (speech), edge models.

Acquisitions (Mistral’s only two to date)

- **Koyeb (Paris; February 2026)** - Mistral’s first-ever acquisition. Serverless cloud infrastructure folded into “Mistral Compute” for ultra-fast AI deployment (targeting sub-200ms startup), auto-scaling, and isolated “Sandboxes” for agentic workflows. Co-founders Yann Léger, Edouard Bonlieu, and Bastien Chatelard joined engineering under CTO Timothée Lacroix.
- **Emmi AI (Linz, Austria; May 2026)** - “Physics AI” for industrial simulations and digital twins (a Universal Physics Transformer running fluid-dynamics and thermodynamics directly on GPUs, without traditional mesh recreation). 30+ researchers/engineers joined the Science & Applied AI teams; co-founder Johannes Brandstetter became VP of AI for Science, and Emmi’s Linz HQ becomes Mistral’s physical-AI hub. Largely share-based, reportedly approaching ~€1bn.
- **Note:** “AnyAI” and “MDI” that appeared in the session recordings were transcription artifacts - Koyeb and Emmi AI are Mistral’s only two acquisitions.

Compute roadmap (across keynote + infrastructure panel)

- Boulogne-Billancourt (Paris) - 40 MW data center, live since early 2026.
- Les Ulis (south of Paris) - 10 MW inference-focused site, live summer 2026.
- Borlänge, Sweden - built through 2027 to host the Vera Rubin generation.
- Greece - next site (closing keynote).
- Stated ambition: ~200 MW by end of 2027 and ~1 GW by 2030.

Partnerships named

- **Industrial / engineering:** Airbus, BMW (crash simulation), Moeve (P&ID), ASML (lithography), EPO (patent document AI).

- **Telecom / hardware:** Ericsson (custom silicon code models), Qualcomm (distributed/edge), NVIDIA & VAST Data (AI-factory infrastructure).
- **Finance:** BNP Paribas, HSBC, ABN AMRO, US hedge funds, Abanca (banking agent).
- **Public sector:** Luxembourg (sovereign AI / legal assistant), EDF (energy / nuclear).
- **Consumer:** Amazon (Alexa+ non-English quality), plus low-resource languages (Darija, Amazigh).

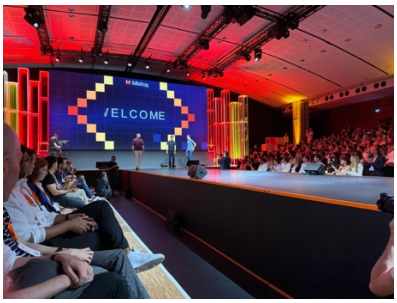
Headline metrics cited

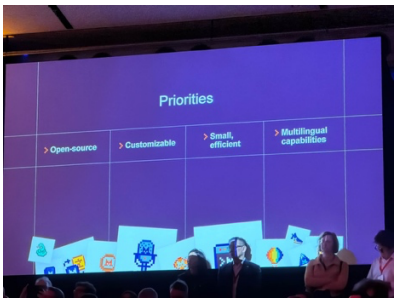
Use case	Result
ASML lithography diagnosis	~120× faster, similar accuracy
BMW crash-sim analysis	30–35 min → ~2 min
BNP Paribas KYC	incomplete files 80% → 10%; weeks → days
Ericsson CBB benchmark	8% → >90%
EPO patent OCR	4× throughput, +21% accuracy, days → seconds
Moeve P&ID	VLM link ID 95% confidence; 94% → target 99%
Abanca "Sophia" agent	95% intent accuracy, 100% task completion, 2M users
Robotics kitting	96% success rate

3. Session-by-session summaries (chronological)

3.1 Opening Keynote - Strategic Pillars, Product Announcements & Partnerships

Arthur Mensch (CEO), Timothée Lacroix (CTO), Guillaume Lample (Chief Scientist)





Arthur Mensch welcomed attendees to Mistral's first conference and set out two strategic pillars: owning the full technology stack (compute and chips through to applications) and deep vertical integration for specific use cases. Timothée Lacroix covered infrastructure and the compute roadmap (Boulogne-Billancourt 40 MW, Les Ulis 10 MW inference, Borlänge/Sweden through 2027) and the Koyeb acquisition. Guillaume Lample outlined model priorities - open-source, customizable, small, efficient, natively multimodal, multilingual (200+ languages) - plus Mistral OCR, agentic search in Medium 3.5, and Mistral Large 4 (summer 2026). Product reveals: the Vibe family (Work + Code), Studio, and Forge. Vertical integration spanned industrial (the Emmi AI physics-AI acquisition, Airbus, BMW), finance (BNP Paribas, HSBC, ABN AMRO), and consumer electronics (Amazon Alexa+). The ASML testimonial (120× faster diagnosis; always-on AI code reviewer) anchored the value story.

3.2 BMW × Mistral - Industrial AI for Crash Testing

Dr. Franz Decker (CIO & SVP Group IT, BMW); Marjorie Genette (CRO, Mistral)

A strategic partnership to co-develop a specialized model for crash-test simulation analysis, trained on BMW's proprietary engineering data. Expected to cut analysis from 30–35 minutes to ~2 minutes while surfacing richer insights. BMW deliberately chose a hard first use case to create a scalable foundation extending to aerodynamics, robotics, and production. Emphasis on embedding AI directly in engineering workflows (a "data flywheel" plus a "people flywheel" as skeptical engineers are won over), with secure on-premise training. Genette stressed "industrial truth" - real, sensitive data - as what makes specialized models outperform general AI.

3.3 Physical AI, Agentic Systems & Robotics

Olivier Duchenne (Robotics Science Lead, Mistral), Pierre Stock (VP Science Ops), Michael Meyer-Ortmanns, + Anna Dolganov (papyrology)

Physical AI uses Vision-Language-Action (VLA) models for generalist machines that navigate, manipulate, and reason. A two-system architecture pairs an agentic VLM for high-level planning (~0.5 Hz) with a low-latency VLA for reactive motor control (5–10 Hz); Robostral 8B posted ERQA 48.4% and RealWorldQA 69.5%. Manipulation highlights: SOTA navigation, zero-shot instruction following, pointing-based selection, and industrial kitting at 96% success, with rapid reprogramming from human demonstration. The agentic segments covered Mistral's portfolio (MM3.5, Voxtral, edge models, Forge) and a training pipeline (domain-adaptive pretraining → SFT with agentic traces → optional online RL), operationalized via harnesses/scaffolds with thinking-trace observability. "Skills" were introduced as evolving, co-authored external memory. A closing segment showcased LLMs unlocking ancient papyrological archives (Austrian Academy of Sciences).

3.4 Luxembourg × Mistral - Sovereign AI & Legal Assistant

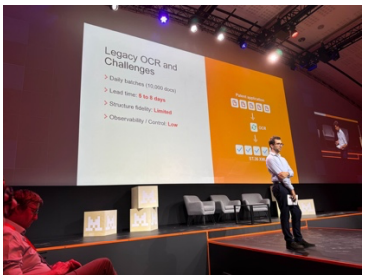
Romain Martin (First Government Adviser, Luxembourg); Kevin Riera (Solution Architect, Mistral); moderated by Guillaume Bour (VP Revenue EMEA)

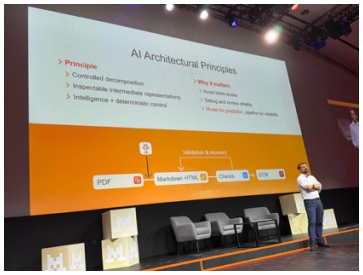
Luxembourg is deploying sovereign AI fully on-premise, managed by local teams on the nation's centralized GPU infrastructure. Sovereignty = local hosting + local control + local expertise. The flagship is a legal AI assistant simplifying access to legal texts for civil servants and citizens, to be extended into a "universal chatbot." Framed as an "entrepreneurial state" leading adoption ahead of the private sector, with the goal of automating compliance (e.g., tax declarations) to cut red tape. Longer-term: Project ESCO for national skills mapping and upskilling, university/research access, and the vision of every civil servant supported by an AI agent. Described as a 5+ year commitment.

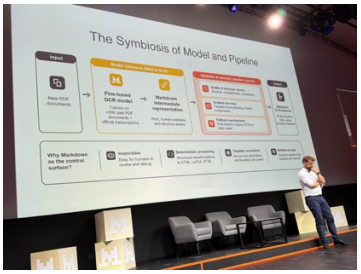
3.5 Panel - AI Factories, Agentic Inference & Deterministic Document AI

Mistral (Brooke Gleason, Yann Léger), NVIDIA (Rod Evans), VAST Data (Alon Horev); + EPO session (Romain Richard, EPO; Maxime Capron, Mistral)

AI infrastructure is now critical infrastructure requiring integrated "AI factories." NVIDIA's "five-layer cake" and rack-based systems (GB200) cited a 10× speedup for a Mistral model; VAST framed data platforms as the key to GPU utilization and the dynamic inference→fine-tuning→RL loop. Mistral detailed owning specialized infrastructure, MOE-based distributed inference, agent sandboxes, and a capacity roadmap (new 10 MW near Paris; ~200 MW by end-2027; ~1 GW by 2030). Heavy emphasis on governance, security, auditing, sovereignty, carrier-grade inference (liquid cooling, disaggregation), and hybrid/federated on-prem-CPU + cloud-GPU patterns. The EPO segment replaced black-box OCR with a fine-tuned Mistral OCR model plus a deterministic pipeline using an inspectable Markdown intermediate - achieving 4× throughput, +21% accuracy, and days→seconds on a single on-prem GPU, with LaTeX for formulas.

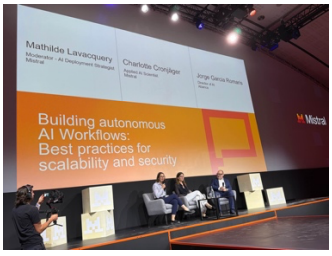


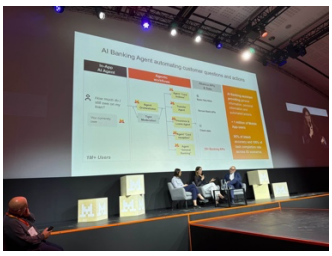




3.6 Abanca × Mistral - Autonomous AI Workflows for Banking

Mathilde Lavacquery & Charlotte Cronjäger (Mistral); Jorge García Romarís (Abanca)



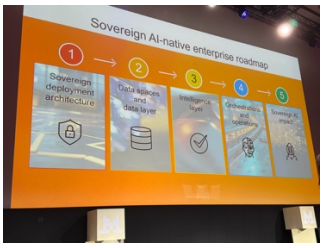


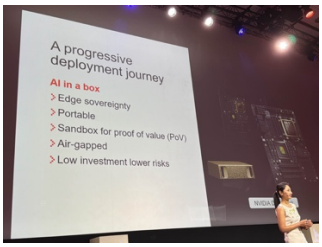
The "Sophia" agent evolved from an informational assistant into a full banking agent across three layers: informational, navigational, and action/transactional. Rolled out progressively to

~2M mobile users; layer preference depends on context, not age/proficiency (stressed users want direct resolution). A central orchestrator routes to specialized sub-agents (cards, loans, mortgages) on Mistral Workflows. Core safety principle: separate the AI's probabilistic outputs from deterministic banking actions via a validation manager enforcing business rules. Multi-layered defenses scan all I/O for PII/toxicity/policy and prompt injection; explorative and adversarial test agents stress the system. Advice: define automated evaluation from day one, start small but architect for scale, and keep AI provider and domain experts close.

3.7 NTT DATA × Mistral - Sovereign AI Partnership

NTT DATA (Amal Tawakuli) & Mistral (Han Héloir)

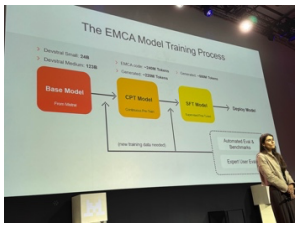




Sovereign AI defined as full end-to-end control of infrastructure, data, and models within an organization's boundaries. NTT DATA (full-stack transformation) + Mistral (models, AI Studio, Le Chat) presented a five-step sovereign roadmap. Deployment patterns: on-premise (LuxProvide supercomputing, OpenShift cluster automation), "AI in a Box" (portable, air-gapped PoV; e.g., NVIDIA DGX Spark), and "AI Factory" (hybrid, production-ready). The data/intelligence layer covered federated learning (AID-Care: Japan dementia data + Luxembourg expertise, exchanging only model weights), a RAG blueprint (keyword/vector/hybrid retrieval + validation + security agents), a teased "Quantum RAG," and a post-quantum cryptography (PQC) readiness agent. An "AI Enterprise Holistic Orchestrator" was positioned as a central automating "brain."

3.8 Ericsson × Mistral - Custom Code LLMs for Proprietary Silicon

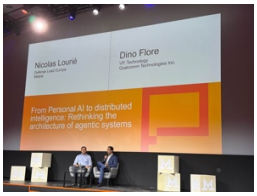
Mistral (Tara Advaney) & Ericsson (Fredrik Ostrom)



Goal: a state-of-the-art LLM that deeply understands Ericsson's proprietary ASICs (which process ~50% of global 5G traffic outside China) to accelerate 5G/6G development and institutionalize expert knowledge. The main challenge was data scarcity - ~80% of effort went to data augmentation/generation. A three-stage training from two base models (Devstral Small 24B, Devstral Medium 123B): continuous pre-training (with an agentic "CodeScribe" generating code descriptions), then SFT on synthetic Q&A. Team: ~15 Ericsson engineers + 5 Mistral experts on Ericsson's AWS (64× H100). Mistral provided base models and the Forge co-training layer (on-prem data stays on-prem). Results (Dec 2025–May 2026): a 24B "ML-developer" model and a 123B model excelling at Ericsson↔x86 code translation; CBB benchmark rose 8% → >90%; now in early production. Next: RL infrastructure for fast PPO simulation loops; clarifying where RLHF is practical.

3.9 Qualcomm × Mistral - Agentic AI & Distributed Computing

Dino Flore (VP Technology, Qualcomm); Nicolas Lourié (Defense Lead Europe, Mistral)



Two simultaneous, non-incremental shifts: AI becoming agentic, and compute moving to a distributed device–edge–cloud model. Power efficiency is the fundamental constraint (from battery devices to data centers), with memory bandwidth second (addressed via near-memory compute). Sophisticated orchestrators dynamically route workloads by device capability, data sensitivity, and latency. Edge model sizes are hardware-bound today: wearables ~1B params, phones ~10B, laptops ~20B, improving roughly every ~18 months. Defense is a prime hybrid use case (mission-critical local processing + deeper secure cloud analysis); a key tension is reconciling 5+ year military planning with ~18-month AI cycles. “Sovereignty” now means controlling compute, data, and control flow across the whole distributed system.

3.10 Moeve × Mistral - Industrial AI Collaboration (P&ID Analyzer)

Ignacio Archondo (Mistral); Iñigo Fernández Sanz (Moeve)

A joint project (started January) converting static Piping & Instrumentation Diagrams (P&IDs) into a queryable graph database. Mistral’s strategy rests on three principles - personalization (fine-tune on engineering IP), sovereignty (client control), orchestration (interconnected platform). Moeve (a 90-year-old Spanish energy company targeting >50% of EBITDA from sustainable activities by 2030) frames “Energy Parks” producing HVO, SAF, and green hydrogen. Technical approach: divide-and-conquer over text/lines/symbols; classify PDFs as vectorized vs rasterized; LLM tag classification beating regex; a VLM identifying cross-plan links at 95% confidence; and an LLM agent translating natural language into SQL/Cypher (e.g., “What is connected to FIC-101?”). Strong internal feedback; accuracy 94% with a target of 99% for production. Next: deployment architecture decision, advanced fine-tuning, and a data-annotation process for unlabeled P&IDs.

3.11 Closing Keynote - “The end of AI as we know it”

Timothée Lacroix (Co-founder & CTO, Mistral)

A recap of the day’s value chain: Compute (own capabilities; new sites in Sweden and Greece; secures capacity and enables full-stack virtualization), Models (text → integrated reasoning + multimodal powering agents/platforms; ongoing research teased), Product/Platforms (Lusha unified under Vibe; agents handling longer tasks once connected to business data), and ROI of AI (tailoring to each client’s unique problems). Named partnerships with financial institutions and industrial firms (Airbus, BMW) and the Emmi AI acquisition for faster physics simulations. Closed by inviting clients to bring their hardest problems.