emma

E-BOOK

AI for SMBs:
Accessible, Affordable, and Unconstrained

IN THIS EBOOK

Table of Contents

CHAPTER 01

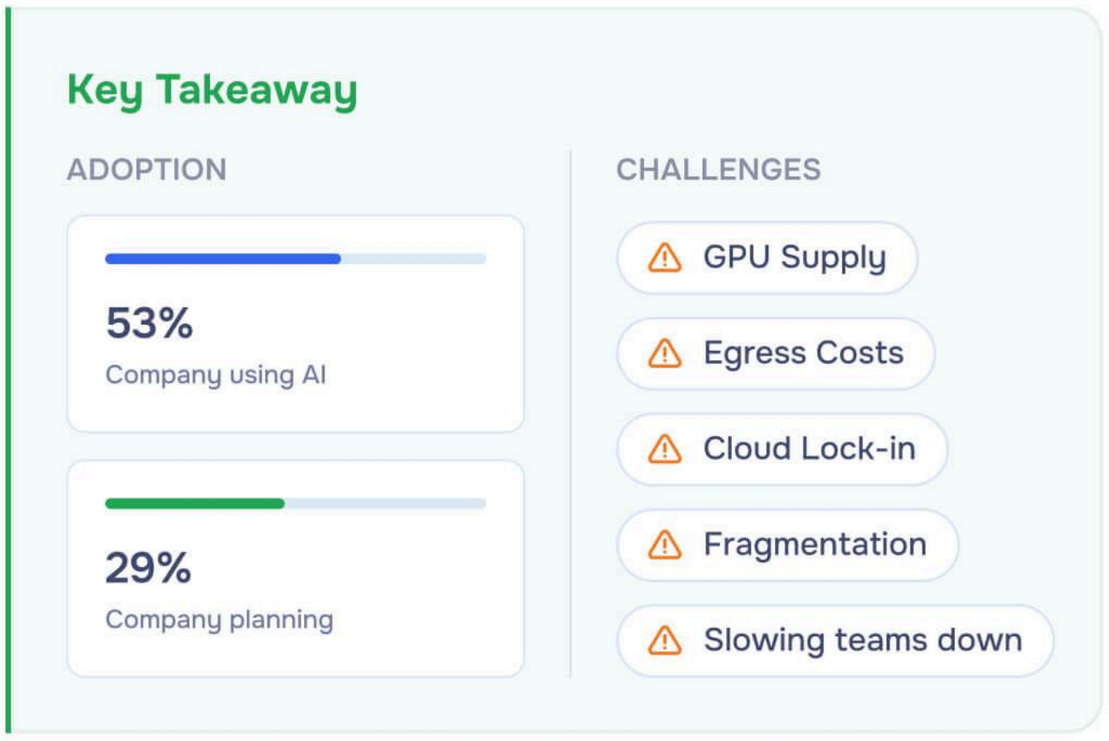
The SMB AI Challenge

Artificial intelligence isn't just for hyperscale data centers or enterprises with massive IT budgets anymore. Today, even small and medium-sized businesses (SMBs) are under pressure to adopt AI to create and deliver AI-powered services — from personalized customer experiences to predictive maintenance, industrial automation, and real-time analytics. According to a report from SMB Group, **53%** of SMBs are already using AI, and another **29%** plan to adopt it within the next year. Out of **18%** that have no plans yet, **smaller firms are the most reluctant to start.**

But getting access to the right infrastructure isn't always easy, and supply, cost, and complexity can quickly slow teams down. Even if a small business has the budget, many high-end GPUs. **Waiting weeks or months for access** can slow AI projects and stall innovation. The **same GPUs might be available in another cloud**, but many AI workloads end up tied to a single provider. That makes it hard to use available capacity elsewhere, optimize costs across clouds, or keep flexibility and negotiating power.

If SMBs do try to move workloads and have adopted interoperability best practices, moving large datasets between clouds or regions becomes another challenge. For AI, where models and datasets can be hundreds of gigabytes or even terabytes, **egress fees can quickly eat up the budget.**

On top of that, fragmented infrastructure adds complexity. When SMBs rely on a mix of on-prem servers, cloud services, and sometimes neo-cloud GPUs, orchestrating training, inference, and storage across all these platforms becomes tricky and error-prone. SMBs usually have fewer people, less operational bandwidth, and less tooling to manage this complexity. So the problem hits them harder than enterprises with ample resources.



emma solves this by giving teams fast access to GPUs across clouds and that too, with control, speed, and flexibility. The platform provides on-demand access to the latest **NVIDIA GPUs, H100, A100, H200, L4**, and more **across AWS, Azure, GCP, Nebius, and emma's Luxembourg-based data center.** Imagine running inference in one cloud, training in another, and orchestrating a logical GPU cluster across multiple providers, without waiting, lock-in, or fragmentation.

CHAPTER 02

Flexible Infrastructure for Every AI Workload

> **To meet the challenges SMBs face, we've structured the emma cloud operations platform to support several infrastructure modalities.**

Multi-Cloud GPU Infrastructure

Instead of being limited by a single cloud's regional quotas or waitlists, emma **aggregates GPU availability across multiple providers and geographies.** If one provider is constrained or backlogged, capacity can be provisioned elsewhere **without requiring new skills, manual configurations, or architectural changes.** This removes one of the biggest bottlenecks in AI infrastructure — waiting weeks for inventory opening up and subsequent approvals.

Teams can request GPUs through emma's governed provisioning flow, scale into NVLink-ready 8-GPU training nodes, or reserve capacity for predictable workloads, all from the same interface with unified billing.

Logical GPU clusters can span several clouds and regions, allowing training and inference workloads to run wherever availability, cost, or end-user proximity is optimal. Combined with emma's private cross-cloud backbone, data movement never becomes a cost barrier, regardless of how data-intensive workloads are.

In most environments, egress is where multi-cloud GPU strategies break down. Even when compute pricing is transparent, egress charges come as a surprise whenever you synchronize model checkpoints between training nodes, replicate datasets across regions, or distribute inference workloads across clouds. Cloud providers have little incentive to make outbound traffic inexpensive, which is why multi-cloud GPU clusters frequently become financially unpredictable. emma largely neutralizes this friction by bypassing the public internet and provider-specific interconnects.

Hybrid Deployments

When AI projects need both elasticity and control, emma lets teams combine multiple infrastructure types into a single, orchestrated environment: cloud GPUs for burst capacity and on-premise, VMware infrastructure for steady-state or regulated workloads.

With hybrid deployments, teams can store, process, and train critical models where data residency and compliance requirements demand it, and burst into the cloud when training demand spikes. They can provision GPU VMs through governed request flow to handle anticipated peak workloads.

Overall, teams can orchestrate a balanced AI architecture, which is flexible, consistent, and fully managed from emma's unified interface.

Managed Kubernetes for GPU Workloads

AI teams that have standardized on Kubernetes face a familiar problem when GPUs enter the picture. Every cloud has its own managed Kubernetes (mk8s) service, its own GPU node pool configuration, its own driver compatibility matrix, and its own cost view. Running GPU workloads on EKS, AKS, and GKE means managing three separate operational models in parallel.

emma's mk8s with GPU support removes that fragmentation. Teams can provision fully managed GPU Kubernetes clusters across AWS, Azure, and GCP from a single interface, in minutes, with NVIDIA driver-optimized, CUDA-validated images ready from the first pod scheduled.

Platform teams define the standard once; data scientists self-serve within guardrails; cost and utilization flow into emma's unified dashboard per project and per team. Supported mk8s GPU types include NVIDIA A100, H100, H200, A10, L4, and T4, covering the range from experimentation to production training and inference. And because clusters are provisioned through emma, they inherit the same RBAC, tagging, and policy controls as every other emma-managed resource.

CHAPTER 03

Engineering AI on a Budget

> **The emma platform gives your teams the levers to match workload to infrastructure, which helps control costs without slowing down innovation.**

Right-sizing GPU Workloads

Different tasks demand different GPU capacity. With emma, teams can instantly find and choose the right infrastructure for the job. Practically, it can look like:

Picking the right GPU for the job avoids overspending while keeping performance intact, as needed.

Observability as the foundation of cost discipline

Right-sizing only works if teams know how the GPUs they've provisioned are actually being used. Without utilization data, allocation decisions are guesswork, and expensive GPUs often sit idle inside training or inference jobs that could run on smaller hardware.

emma's built-in GPU Monitoring closes that gap. At the VM level, GPU Utilization, vRAM Usage, and vRAM Utilization show whether compute and memory are being worked or whether an H100 is spending most of its cycles waiting on data. At the Kubernetes cluster level, GPU metrics in the mk8s monitoring tab extend the same visibility to node-level workloads, including power and temperature indicators that surface thermal throttling before it silently degrades training throughput.

Burst vs steady-state modeling

emma makes it virtually seamless to keep predictable, steady workloads on reserved GPUs and request on-demand cloud GPUs only when demand spikes. This way, surplus capacity is always available when needed, without paying for idle GPUs.

Transitioning between environments is seamless. No extra skills, scripts, or complex coordination required. Teams manage everything from a single control plane, where all data, usage stats, and monitoring are centralized. Scaling up or down is simple, giving SMBs predictable performance and cost control without operational headaches.

Standardizing inference to contain deployment cost

Inference is often the place where AI infrastructure costs quietly escalate. Each team writes their own setup scripts, picks their own instance sizes, and installs their own runtimes, meaning every deployment is a fresh opportunity to overprovision, pick the wrong GPU, or skip cost controls entirely.

emma's Inference Workflows let platform teams define reusable, governed templates with VM size limits, parameter constraints, and RBAC controls baked in. Application teams deploy from an approved library in minutes, within guardrails the platform team set once. Every deployment shows a cost preview before it launches, and deployed instances behave like standard emma VMs, with GPU monitoring available and costs attributable to the team that deployed them.

Data locality control

In modern AI workloads, data gravity often outweighs model size. A trained model might be a few GBs, but the datasets behind it can be hundreds of GBs, terabytes, or even petabytes. However, GPU capacity may not always be available in the same location where data resides. And moving that data across regions quickly becomes the real bottleneck.

With emma, teams can provision GPUs from other cloud providers that have availability **within the same region where the data already resides.** Instead of relocating datasets, compute is brought closer to the data.

Combined with emma's optimized private inter-cloud networking backbone, this setup minimizes egress costs, reduces latency, and helps preserve regulatory boundaries when data is already stored in compliant zones.

Multi-cloud arbitrage

With emma, teams can deploy workloads wherever required GPUs are available and pricing is optimal, without changing tooling, rewriting infrastructure code, or rearchitecting their stack. Instead of being locked into a single provider's pricing model, SMBs can build logical GPU clusters that span multiple clouds, and they can shift workloads when market conditions change. If GPU prices spike in one region, they can move workloads to a more cost-efficient provider or region. If capacity tightens, additional GPUs can be provisioned elsewhere and orchestrated as part of the same distributed environment.

Because every environment runs through the same orchestration and networking layer, deployments feel identical across providers. The control plane remains the same, the management experience remains the same, and operational overhead doesn't increase.

CHAPTER 04

Sovereign AI on a Budget

> **For European SMBs handling sensitive data or operating in regulated industries, compliance requirements often introduce additional infrastructure constraints.**

Maintaining strict data residency or sovereign control doesn't just limit regions where workloads can run, it also restricts which cloud providers can be used. That often means reduced flexibility to tap into hyperscaler-grade services, global GPU availability, or advanced AI tooling. It creates a tradeoff between sovereignty and innovation.

emma, being a European company under Luxembourgish jurisdiction, offers a unique advantage for sovereignty-bound organizations. emma offers powerful infrastructure for EU-hosted workloads in its Luxembourg-based data center. Whereas, emma's roadmap includes adding GPU instances from EU-native providers such as IONOS, OVHcloud, and Gcore. It gives the flexibility to orchestrate sensitive data and regulated workloads on EU-native infrastructure, while allowing non-sensitive training or inference workloads to scale across hyperscalers and other providers where GPU pricing or availability is more favorable.

emma powers several architectural capabilities that make sovereignty economically viable:

- **Multi-cloud deployments** keep sensitive data on EU-based sovereign infrastructure , while hosting non-sensitive workloads on best-price cloud GPUs.
- **Optimized private networking** reduces egress charges in hybrid and multi-cloud AI deployments.
- **Centralized governance and orchestration** eliminate the need for separate management stacks per environment, saving time and effort.

Because workloads can be distributed instead of confined to a single sovereign environment, organizations avoid the common cost traps of sovereignty. SMBs can meet regulatory obligations without isolating themselves from innovation. This makes sovereignty operational, not just a legal constraint.

CHAPTER 05

Accessible, Affordable AI, Anywhere

> **AI becomes truly accessible when infrastructure isn't locked into a single provider, pricing model, or deployment type.**

By combining multi-cloud and hybrid options, teams can choose the right environment for every workload, whether that's any neo-cloud with available capacity, or hyperscalers for the latest generation GPUs.

This flexibility directly impacts cost and availability. Instead of overpaying for a hyperscaler region or waiting on limited GPU supply, teams can distribute workloads wherever demands are met.

However, managing multiple infrastructure environments can get complicated fast. Each has their own setup, provisioning, and orchestration requirements. Without a unified approach, teams spend more time wrangling infrastructure than building AI. At present, 78% of engineering groups say they wait a day or more for infrastructure support from SRE/DevOps teams, and 16% report waits of a week or longer.

emma combines various AI resources into a single, orchestrated environment behind one powerful, intuitive dashboard suitable for every stakeholder — infrastructure teams, platform engineers, administrators, and finance.

Source: 2025 State of internal developer portals

Platform Benefits

- **Unified management:** All cloud CPUs and GPUs are visible, controllable, and billed in one place.
- **Fast deployment:** Hybrid environments can be set up and expanded quickly, without complex manual integration.
- **Cost optimization:** Match infrastructure and configurations to precise workload needs, minimizing idle or overprovisioned resources.
- **Growth-ready:** Easily scale deployments across environments as AI projects grow, without re-architecting pipelines.

TAKE THE NEXT STEP

Ready to scale AI without the trade-offs?

AI infrastructure doesn't have to be expensive, fragmented, or locked into a single provider. With platforms like emma, organizations gain the flexibility to run workloads wherever it makes the most operational and financial sense, across multiple clouds, and hybrid environments.

With a unified orchestration and management layer, even SMBs with lean teams can strategically design architectures that balance compliance, innovation, and affordability without adding operational complexity. Instead of being forced into tradeoffs, they gain the control to meet regulatory requirements, the freedom to tap into hyperscaler-grade innovation, and the efficiency to keep costs predictable and optimized.

$300K in cloud credits

For teams actively building and preparing production-ready AI workloads, emma is currently offering access to up to $300K in cloud credits to help accelerate real deployments — no complicated hoops, no hidden strings. If you're ready to scale something real, apply and put the credits to work.

APPLY FOR CREDITS →

Explore the demo portal

Spin up multi-cloud GPUs yourself in the demo portal and experience the deployment workflow hands-on — no sales calls, no pressure, just infrastructure at your fingertips.

TRY THE PLATFORM →

About emma

The cloud operations platform for distributed infrastructure

emma is built for organizations running distributed workloads across hybrid and multi-cloud environments — where operational complexity, fragmented tooling, and unpredictable costs get in the way of building.

It provides a single policy-driven operating layer that spans hyperscalers, regional European cloud providers, AI-optimized infrastructure, and on-premises environments. It gives engineering, platform, and finance teams unified deployment, governance, policy and cost control from one place. Sovereignty and compliance are built in through proactive guardrails, not bolted on.

With a vendor-neutral architecture, emma ensures organizations can choose the right infrastructure for cost, performance, and regulatory requirements — without lock-in or operational trade-offs.

Unified Operations

The entire distributed stack — hyperscalers, sovereign clouds, AI providers, on-premises — operated from a single interface. No context switching. No blind spots.

Integrated Networking

emma's private networking backbone is built for data-intensive multi-cloud operations. When data moves over emma's private networking backbone, egress costs become predictable compared to the public internet.

Automated Governance

The same policies, governance, and deployment logic apply across every environment. Compliance holds everywhere, by design.

Extensible by Design

New providers are added without rebuilding how everything is operated. emma's operating layer absorbs the new environment, while existing governance, cost model, and deployment patterns stay intact.

Intelligent Workload Orchestration

Workloads are placed across providers based on performance, economics, and policy — not vendor incentives. No commercial bias toward any hyperscaler, neocloud, or AI provider.

Continuous Cost & Performance Optimization

Unified visibility into cloud usage, performance, and cost across every environment. Workload placement decisions that optimise economics without compromising governance.

2021
Founded & HQ in Luxembourg

15+
Cloud providers supported

~90
Cloud engineers at your service

NO
US Cloud Act applicability

CERTIFICATIONS & FRAMEWORKS

emma operates under internationally recognized security and compliance frameworks, including ISO-certified security management and SOC 2 Type II audited controls, with data protection aligned to GDPR and resilience aligned with NIS2 and DORA.

NEXT STEPS

Contact us

emma.ms

info@emma.ms

19-21 Rte d'Arlon, 8009 Strassen, Luxembourg

emma | Make Cloud Work for You