



Put GenAI to work

A startup's practical guide to scalable,
secure, and cost-effective AI



The startup dilemma

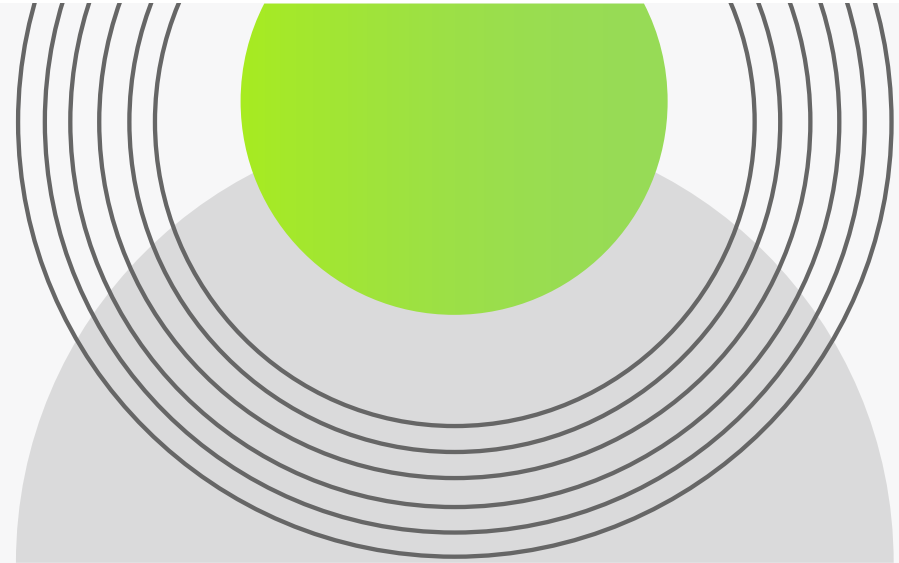
Generative AI is no longer experimental—it's embedded in the product roadmap of nearly every ambitious startup. From building smart assistants to automating customer workflows, GenAI offers a path to faster releases, deeper personalization, and new revenue models.

But the leap from prototype to production-grade GenAI is harder than it looks. Many teams hit the same walls: inconsistent output, unpredictable costs, poor user trust, and fragile infrastructure.

30%

of Generative AI projects will be abandoned after proof of concept by end of 2025.¹

1. Gartner, 2024, <https://www.gartner.com/en/newsroom/press-releases/2024-07-29-gartner-predicts-30-percent-of-generative-ai-projects-will-be-abandoned-after-proof-of-concept-by-end-of-2025>



Whether you're experimenting with RAG pipelines, exploring agentic AI, or standing up a LangGraph-based orchestration layer, the challenges are real:

- How do you serve models reliably at scale?
- How do you govern and secure evolving knowledge bases?
- How do you control spend—and build observability into LLM workflows from day one?

This is the startup dilemma. The same tech that can define your next funding round can also balloon your cloud bill – or stall at the compliance gate.

To win, it's not enough to build with GenAI. You must operationalize, secure, and scale it—fast.

This is where Automat-it and AWS can help you.

What's holding back GenAI innovation?

Startups are powered by innovative ideas. But the time, skills, and robust infrastructure required to realize them remain at a premium. Working with a trusted service provider with extensive experience gives you the scope, people power, and hours in the day to turn ideas into outcomes and scale successfully.

At Automat-it, while working in this space we have identified five consistent barriers that stand in the way of GenAI projects:

1. Unclear ROI – Business leaders struggle to connect GenAI investments to specific and measurable business outcomes.

2. Infrastructure inefficiencies – Without AI-optimized environments to work in, the chance of underperformance or escalatory spending is high.

3. Rising cloud costs – Training and inference expenses can spiral out of control without the right FinOps oversight and discipline.

4. LLMOps complexity – Traditional deployment pipelines can buckle under escalating demands of model experimentation and iteration.

5. Compliance and risk – Concerns around hallucinations, data leaks, and governance gaps lead to hesitation, slowing down adoption and time to value.

only 25%

of executives report seeing significant value from their AI initiatives, according to Boston Consulting Group (BCG)²

The challenge is real—but these barriers can be overcome. Success depends on having a partner who not only understands the GenAI technology stack, but also how to deploy and scale it in the real world. Automat-it brings deep experience working with startups, helping you to move faster with proven infrastructure templates and deployment patterns already used by your peers.

Real business benefits of working with Automat-it:

- Faster time to value
- Fewer false starts
- Faster pathway from prototype to production

2. Boston Consulting Group. Closing the AI Value Gap: 2025 AI Radar. February 2025. Available at: <https://www.bcg.com/publications/2025/closing-the-ai-impact-gap>

Building a GenAI startup stack that simply works

Startups that succeed with GenAI aren't always those with the biggest models, but those with the smartest execution. There are a few core principles we see in GenAI projects that deliver value and effectively scale from pilot to production.

- ✓ Lean, cloud-native systems
- ✓ DevOps discipline
- ✓ Robust business alignment

Succeeding with scalable GenAI projects requires more than the basics of code and compute. You must build or employ systems that are:



Infrastructure-aware:
Architected for performance, scale, and flexibility



FinOps-optimized:
Governed with visibility and cost controls



Business-integrated:
Tied to real workflows and measurable outcomes



Secure and compliant:
Built with trust and explainability in mind



DevOps-aligned:
Repeatable, testable, and deployable via CI/CD

Next, we'll show how you can realize these goals and create GenAI success at scale >



Achieve GenAI success with the help of **Automat-it and AWS**

Automat-it is an AWS Premier Partner with deep GenAI expertise and a track record of helping startups move fast—without cutting corners. Our approach combines AWS’s cutting-edge tools with hands-on experience and our tried-and-tested infrastructure templates to turn complexity into competitive advantage.

Enjoy speed without compromise

Automat-it delivers lean, production-ready GenAI pipelines that match startup velocity. We reduce time to production by leveraging reusable infrastructure templates—so you don’t have to start from scratch. Services such as Amazon SageMaker JumpStart and the [Bedrock Migration Accelerator](#) help you move fast, with a strong technical foundation from day one.

We also help you go to market faster through the [AWS Marketplace Accelerator Suite](#), enabling seamless onboarding to AWS Marketplace so you can unlock visibility, reach new customers, and deploy with confidence.

Escape common pitfalls

From runaway cloud costs to model underperformance, the risks of GenAI are real. Our tailored infrastructure is designed for your use case, with FinOps built in from day one. We use Umbrella to identify cost-saving opportunities, we monitor GPU spend in real time, simulate Bedrock deployment costs, and surface early warnings before budgets are blown. Deployment guardrails also catch issues such as GPU overprovisioning or untested model releases before they impact your business.

Achieve GenAI success with the help of **Automat-it and AWS**

Gain visibility where it matters

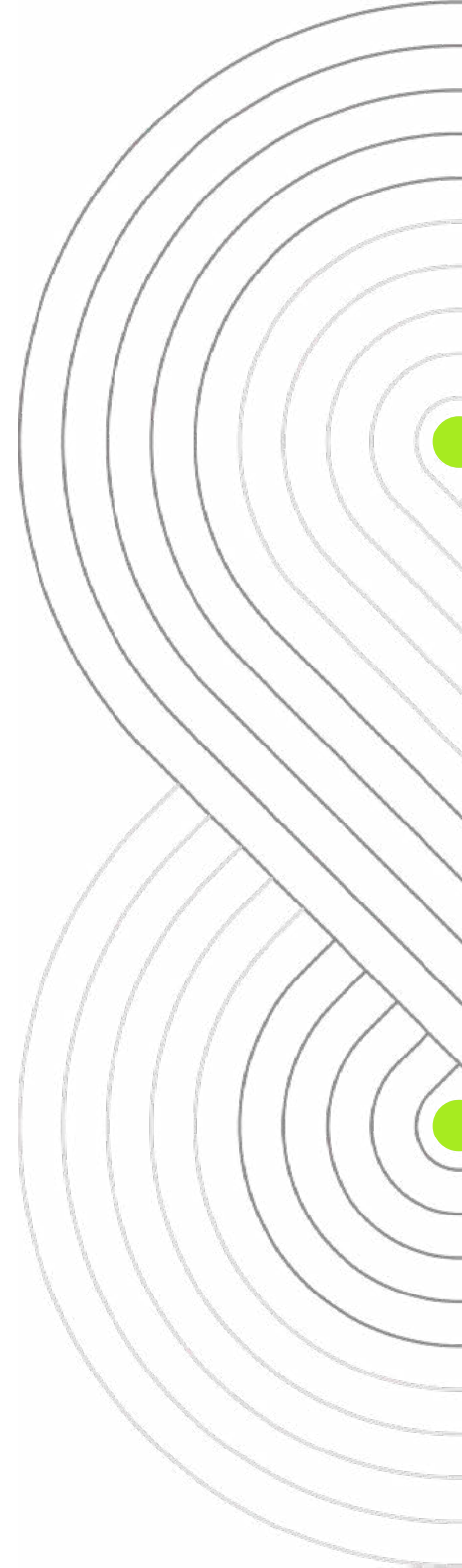
Deploying a model is just the beginning. We embed observability into every solution, tracking latency, cost drift, and model behavior in real time. Tools such as Langfuse and custom FinOps dashboards turn model performance into actionable insights—helping you iterate, optimize, and extract value over time.

Build with trust as standard

Security and safety are core to our GenAI approach. From input/output validation to secure defaults and LLM-specific protections, we embed trust—and the necessary guardrails like responsible model behavior—from day one. That means safer systems, fewer surprises, and peace of mind for you and your users so you can operate with confidence.

Start with tailored foundations

Great GenAI outcomes start with the right architecture. We lead hands-on ideation sessions to deeply understand your goals and design aligned, scalable systems—from LLM choice to inference stack. Architecture diagrams define everything upfront, including prompt engineering flows, RAG pipelines, and governance layers. This clarity avoids missteps, rebuilds, and wasted time later on.





See start-up success in action

Our approach is not purely theoretical. We've applied it for startups across multiple industries to help build, fine-tune, and scale GenAI solutions that produce real improvements in performance and growth.

Continue reading to get inspired by what's possible.



Optimizing foundational model training | Bria.ai

Bria uses GenAI to help teams create images at scale. To improve the efficiency of training their proprietary text-to-image foundation models, they needed to optimize performance and control costs on AWS.

Automat—it's MLOps team supported Bria in integrating the MosaicML Composer framework into their AWS-based training pipeline. We developed a custom, containerized setup tailored to Bria's Visual Generative AI models, unlocking more efficient GPU usage and enabling advanced training strategies.

[Read the full story](#)

40% reduction in training costs

57% lower training time



Faster iteration cycles to accelerate scaling of future workloads



Delivering GenAI that works, fast | Toasty

Toasty, an eco-responsible e-commerce platform startup, needed to resolve inefficiencies in the manual process of updating product descriptions and scraping customer reviews from brand websites.

Automat—it developed a Proof of Concept (PoC) based on Amazon Bedrock that has since moved into production, to automate these processes while staying in line with Toasty's editorial style. The GenAI solution references existing descriptions and applies category-specific rules to ensure consistent output. Integrated with AWS Amplify, it ensures secure, stable, and scalable operations, to eliminate manual scraping tasks and increase accuracy.

[Read the full story](#)

<1 month to train and implement a GenAI solution

70–90% error-free sessions through reduced script errors

2x faster data enrichment for operational teams

Find yourself on the GenAI map

When it comes to Generative AI, knowing where you are is the first step to unlocking what's next. Whether you're still exploring or already scaling, Automat-it can meet you where you are—and help move you forward faster.



Let's turn your GenAI vision into scalable success

You've seen what it takes to realize the possibilities of GenAI beyond experimentation. Now it's time to act.

Automat-it empowers fast-paced startups to grow, deliver and win with strategic, hands-on cloud DevOps, FinOps, and GenAI expertise. With us, you can harness the possibilities of the cloud during every phase of your growth.

Talk to Automat-it today and together we can turn your GenAI ambitions into a scalable, secure, production-grade reality.

[Book a discovery call](#)

