Krista™

The Illusion of Simplicity of Building Your Own Generative AI Systems

Table of Contents

Introduction

Every business must decide whether or not to deploy Generative AI (GenAI). The real question isn't if but how. Building GenAI systems in-house looks deceptively simple, but it's a recipe for wasted budgets, delayed results, and falling hopelessly behind the competition.

The uncomfortable truth is that traditional software development processes just won't cut it. AI projects aren't just 'more complex'—they're fundamentally different. Public tools like ChatGPT and Gemini make it seem easy, but they mask an iceberg of hidden challenges. Think of them as the 'tip'—the real mass lies beneath the surface.

It's time to confront the hard reality. Approaching GenAI using your traditional SDLC approaches will drain your IT teams, devour your resources, and distract you from the strategic innovation your business desperately needs.

Krista is the solution. It's the platform built for the express purpose of enterprise-grade GenAI. Krista isn't another tool – she's the automation engine that lets you build AI solutions with unmatched speed, flexibility, and security. The choice is clear: struggle with building and assembling open-source and point solutions or move at the speed of AI with Krista.

Building Your Own GenAI Platform Appeal (and its Pitfalls)

The initial attraction of assembling hardware and software to deploy GenAI solutions is understandable. Experimenting with public AI interfaces like ChatGPT and Gemini inspires visions of integrating similar capabilities in-house. However, assembling or building your own platform is deceptively complex.[1]

Generative AI is Only 5% of the Solution

Beyond each large language model (LLM) lies a complex web of tasks vital for a successful enterprise GenAI deployment. Ingesting content from diverse sources like PDFs, Word documents, Excel files, apps, and systems presents significant challenges due to differing formats and structures. Ensuring this content is properly prepared can be surprisingly labor-intensive, requiring specialized technology and human oversight. Integrating LLMs with existing systems to access real-time data flows is another major hurdle. GenAI solutions must query and

retrieve data from live information sources to provide up-to-date and contextually relevant answers.[2] This involves sophisticated development work and a deep understanding of both legacy systems and modern AI interfaces. Security and privacy become paramount as sensitive business information is entrusted to GenAI. Encrypting data, controlling access, and constantly monitoring for vulnerabilities are all non-negotiable tasks that demand dedicated resources. Finally, there's the continuous development and maintenance costs; models must be refined, integrations updated, and the entire system must be scaled to meet growing needs. For IT leaders, this translates into a project trajectory that is far more complex, costly, and time-consuming than initial impressions from public AI interfaces suggest. Attempting to build this infrastructure in-house can quickly divert teams away from core strategic initiatives and strain budgets.

Open-Source Illusion and Required AI Skills

Open-source tools like Lang Chain may promise a low-cost entry point, but they quickly expose hidden expenses and complexities. Don't be fooled by the "free" label – successful deployment demands a high degree of specialized AI engineering expertise. Customizing these tools to match your unique business needs rarely involves plug-and-play simplicity. Integrating open-source components into your existing enterprise systems requires extensive development work and deep knowledge of both AI and legacy infrastructure. And the rapid evolution of open-source AI software means you're in for a never-ending cycle of compatibility updates, security patching, and performance optimization. The seemingly low-cost solution morphs into a major resource drain. For business and IT leaders focused on true automation and AI-powered transformation, open-source approaches present the illusion of savings. In reality, they demand specialized AI talent (which is in short supply and increasingly expensive), introduce unpredictable development timelines, and risk unforeseen compatibility issues that disrupt strategic plans and delay ROI.

Specific AI skills you'll need

Building and maintaining an effective enterprise GenAI solution demands specialized knowledge and skills that most traditional software development teams simply don't have. Here are some of the crucial AI skillsets involved:

- **Machine Learning (ML) and AI Capabilities:** It is crucial to understand different AI algorithms and capabilities (beyond LLMs), optimization techniques, and process orchestration.
- **Natural Language Processing (NLP):** GenAI requires handling the complexities of human language, from tokenization and stemming to sentiment analysis.

- **Data Engineering and MLOps:** Preparing and managing diverse data, deploying models, and monitoring their performance as they scale requires specialized skills.
- **Prompt Engineering:** Crafting effective prompts is both an art and a science, and it directly impacts the quality and relevance of outputs from LLMs.[3]
- **AI Ethics and Responsible AI:** Understanding biases, ensuring fairness, and building explainable models are non-negotiable for enterprise success.

Innovation Lock-In

AI and the broader GenAI landscape are evolving at an unprecedented pace. This means choosing to build and maintain a bespoke GenAI system based on a single LLM risks locking your organization into an exhausting cycle of continuous updates, upgrades, and adjustments. Merely keeping up with the latest language model improvements becomes a major strain on resources. With major public LLMs like ChatGPT and Gemini releasing dozens of significant updates in a matter of months, dedicating in-house teams to test, evaluate, and integrate these constant changes is a major undertaking.[4] Failure to do so could result in spiraling costs as performance degrades on older models or risk your GenAI solution lagging behind competitors leveraging the latest advancements.

This "innovation treadmill" is particularly risky in the context of automation. Even subtle changes in LLM behavior can disrupt carefully designed workflows and processes. Imagine your painstakingly crafted automation suddenly starts misinterpreting dates or tables due to a model update. This could lead to wrong decisions, broken processes, wasted time, and erode confidence and trust. For senior business and IT leaders, innovation lock-in presents a double-edged sword. Not only does it divert resources away from strategic initiatives, but it also threatens to render your AI-powered automations stale and potentially even less efficient than traditional processes due to a lack of continuous optimization.

Beyond the Surface: The Hidden Maze of Generative AI Complexity

Content Ingestion

The seemingly straightforward task of feeding documents into GenAI systems is where the complexity begins. While PDFs, Word documents, and Excel files are cornerstones of modern business, their diverse structures, embedded images, and complex tables make them notoriously difficult for language models to process accurately.[5] Naively uploading these files to

an LLM often results in garbled content or poor results from the GenAI assistant. The old adage, "garbage in, garbage out," also applies to this process. Ensuring proper ingestion requires specialized tools to handle these different formats and often involves manual supervision to maintain quality. Additionally, many businesses have accumulated vast content repositories in legacy formats, further compounding the issue. Content ingestion challenges will directly impact the viability and accuracy of your AI-powered projects, so take special care in understanding your different content formats. Attempts to automate processes that rely on poorly ingested data can lead to a cascade of errors, ultimately creating more work or critical errors rather than streamlining operations. The hidden labor needed to refine content ingestion and the ongoing maintenance it requires as new document types are introduced can derail project timelines and drain budgets intended for true innovation.

Real-Time Data

Effectively integrating static content with real-time data sources is crucial for GenAI to provide accurate and timely responses. The most powerful applications go far beyond a search engine for existing documents. To truly transform business operations through automation, your GenAI solution needs to integrate into real-time information flows.[2] This could include anything from live inventory levels and customer order statuses to open contact center issues and accounting systems. Integrating these diverse data sources, each with its own unique structures, protocols, and access rules, often involves complex custom integration and development work. Moreover, ensuring that this integrated information is up-to-date and presented to the GenAI in an easily understandable way presents further development challenges. Preparing your different data sources significantly expands project scope as teams grapple with legacy systems, APIs, and data governance protocols to deal with AI systems that most don't yet fully understand. Automation initiatives that hinge on this real-time data integration can stall if the AI system can't access and correctly interpret the information it needs to make informed decisions and generate content.

Data Privacy

The power of GenAI comes with significant privacy and compliance challenges. While traditional security is vital, it's merely the starting point. LLMs pose unique risks due to their ability to capture, store, and potentially reuse sensitive information. Imagine feeding confidential contract details, intellectual property, or personal data like social security numbers into an LLM – that information could linger within the model, be accessible to others, or be used in unexpected ways in the future. Trying to scrub sensitive data from an LLM is like removing a single drop of water from an ocean without knowing its location.[6] Compounding this issue, regulations like GDPR and industry-specific standards impose strict controls on how data can be used, stored, and shared. With LLMs lacking the ability to selectively delete or "unlearn" data, even basic compliance, like responding to an individual's "right to be forgotten," becomes a major hurdle. It is imperative you prevent sensitive data from ever entering an LLM while ensuring compliance. Avoiding LLMs is not the answer. Your business will not survive. Your only choice is to use a

platform to help you maintain compliance and privacy using the right security and access controls.

Performance & Scalability

Ensuring accuracy, speed, and consistent results as usage of your AI solution grows puts immense pressure on infrastructure and development teams. The demands of a successful enterprise system are far more complex than what is seen when experimenting with hosted public platforms running in enormous data centers. Scaling AI infrastructure, especially with specialized hardware and complex networking requirements, can be costly, even with open-sourced models, and predicting usage patterns and provisioning capacity accordingly is crucial to prevent sluggish performance that can frustrate users.[7] Moreover, the need for any AI system to handle increasing concurrent requests amplifies potential bottlenecks in your code or system design. You shouldn't assume you can model the infrastructure requirements based on your previous per-user software development projects. AI is much more challenging to scale and could lead to unexpected expenses related to hardware and emergency development work.[8]

The Maintenance Beast

The ongoing cycle of updates, bug fixes, testing, and re-tuning models to maintain relevance can quickly drain resources (if you can hire them) and distract from innovation. The growing cost of maintaining bespoke software resonates strongly here. For example, OpenAI's ChatGPT has had twenty-nine major releases from December 2022 to February 2024, and that pace won't slow.[4] AI solutions are not "set it and forget it" propositions. Your AI system must adapt as models and internal processes evolve, new data sources emerge, or the competitive landscape shifts. This requires continuous monitoring, testing, and development work to keep it aligned with your business goals. The cost of internal development teams dedicated to maintaining this system is often grossly underestimated, echoing the lessons learned from decades of custom software projects where maintenance eventually consumes the majority of the budget. For IT leaders, the "maintenance beast" siphons off resources originally intended for innovation and strategic initiatives. Teams become bogged down with firefighting unexpected compatibility issues or performance regressions, hampering their ability to deliver new AI-powered automation projects that drive growth. Moreover, the opportunity cost of delayed innovation can be significant, as competitors potentially pull ahead by adopting solutions that require less internal maintenance overhead.

The Economic Argument: Why Building Your Own GenAI Platform is a Costly Gamble

The transformative potential of Generative AI is undeniable. However, embarking on assembling or developing your own enterprise-grade AI deployment platform comes with significant hidden

costs that can undermine the very benefits you're seeking. While the initial appeal of self-hosted or public-only solutions might seem financially attractive, they often lead to unpredictable expenses, security vulnerabilities, and missed opportunities.

The Price of Complexity

Remember, your GenAI project's success is not just about the LLM model. Content ingestion, real-time data integration, privacy and compliance, and constant maintenance translate directly into costs. Here's why developing your own platform leads to spiraling expenses:

- **Public AI Services:** Per-interaction fees seem attractive initially but can skyrocket with usage, especially if frequent error correction is needed. Unpredictable expenses make budgeting and ROI tracking a nightmare.
- **Self-Hosting Your Own LLM**:
    - **Specialized AI Hardware:** A single server capable of running even a small open-source model can cost upwards of $209,000 or $5000 a month for hosting services.[8] You'll need additional servers for resiliency and several pre-production environments for development and testing.
    - **Talent Acquisition:** Finding, hiring, and retaining dedicated AI talent commands top salaries and can further drain budgets and extend your project timelines.
    - **Protracted Development:** Building a viable GenAI solution from scratch takes months or years, requiring skills you most likely don't have today. This delays potential returns and compounds the financial burden.

Hidden Costs of DIY Solutions

Both self-hosted and public-only approaches mask significant indirect expenses:

- **Unreliable Outputs:** Inconsistent or error-prone GenAI results lead to rework, resource waste, and user frustration. This means unexpected labor and development costs to keep your AI solution functional.
- **Ongoing Maintenance:** Security patches, model updates, performance optimization, and prompt re-engineering are never-ending tasks, especially with self-hosted deployments. This requires a dedicated team, adding to your ongoing expenses.
- **Opportunity Costs:** Drawn-out development cycles and infrastructure battles prevent your IT teams from focusing on strategic initiatives. The potential value GenAI could bring to other business problems remains untapped, representing a major lost opportunity.

The Bottom Line: Building your own enterprise GenAI platform is fraught with financial and opportunity risks. Unpredictable costs, potential security vulnerabilities, and delayed ROI can ultimately jeopardize this technology's transformative potential for your business.

Krista: The Path to Financial Success and Security with GenAI

Krista's platform approach directly addresses the aforementioned economic challenges, making it the smart financial choice for enterprise-wide GenAI. Choose Krista for:

- **Predictable Costs:** Krista's transparent pricing model ensures control and makes budgeting your AI projects straightforward.
- **Accelerated Time-to-Value:** Krista streamlines deployment, allowing you to realize returns on your GenAI investment exponentially faster than in-house-built solutions.
- **Efficiency at Scale:** Krista handles the core, complex GenAI tasks like prompt engineering, enabling your teams to focus on strategic automation and value creation – not infrastructure wrangling.
- **Future-Proof Investment:** Krista evolves with the shifting GenAI ecosystem, minimizing obsolescence risk and protecting your long-term investment. Changing AI services or models in Krista is a click of a button.

Let Krista Handle the Complexity So You Can Reap the Rewards

By shifting the cost equation to your advantage, Krista empowers you to realize the full transformative potential of GenAI at unprecedented speed without the financial uncertainty and technical burdens of in-house development.

Krista as the Game Changer: Accelerating Enterprise-Grade GenAI Deployment

Generative AI promises to transform nearly every electronic business process, but the path from experimentation to enterprise-grade implementation is rife with complexity. While public AI interfaces offer a glimpse of the possibilities, the reality of building and maintaining in-house GenAI solutions presents challenges that can derail projects and delay the realization of strategic benefits. Krista is a purpose-built platform that eliminates the roadblocks hindering

GenAI efforts, allowing organizations to deploy secure, scalable, and effective automations rapidly. Here's how:

Mastering Enterprise Content Ingestion

Krista's sophisticated content ingestion engine is designed to handle the diverse document formats commonly found in enterprise environments. She processes PDFs, Word documents, and Excel spreadsheets, prioritizing accuracy and structural integrity. By minimizing errors and ensuring data quality from the start, Krista reduces rework and delivers reliable results to limit the pitfalls of the "garbage in, garbage out" phenomenon.

Harnessing the Power of Real-Time Data

Krista empowers businesses to leverage their existing information assets through robust integration capabilities. She connects directly with enterprise systems like CRMs, ERPs, and support systems, bypassing the need for time-consuming custom connector development. This allows GenAI solutions to access and process real-time data, enabling powerful business process automations that drive efficiency across operations.[2]

Prioritizing Privacy and Compliance

Unlike public LLMs, which offer limited control over sensitive data, Krista provides the tools to ensure granular data governance and security. Administrators have full visibility into data storage, access, and retention, empowering compliance with mandates and regulations like GDPR and risk management policies. This reduces risk and builds trust, paving the way for utilizing GenAI in even the most sensitive use cases.

Rapid Deployment and Reduced Overhead

Krista eliminates the need for in-house AI infrastructure development and specialized talent acquisition. Organizations can focus on defining and deploying AI strategies, not wrestling with servers or complex system configurations. Krista accelerates time-to-value, transitioning projects from concept to production in a fraction of the time required for custom software development approaches.

Empowering Users and Adapting to Change

Krista's intuitive conversational interface lowers the technical barrier to entry. Business users can interact with and benefit from AI applications without requiring any AI expertise. Additionally, Krista is designed to avoid innovation lock-in and evolve with the AI landscape, offering flexibility for future customization and integration of new LLMs as your AI ambitions grow and scale.

The Clear Choice for Enterprise GenAI

Attempts to build and maintain GenAI solutions in-house can lead to cost overruns, security vulnerabilities, project delays, and user frustration. Krista is the purpose-built platform that mitigates these risks. She accelerates the deployment of accurate, secure, scalable, and adaptable AI solutions, allowing businesses to focus on innovation, not infrastructure. With Krista, organizations can harness the transformative power of GenAI and achieve tangible ROI at unprecedented speed.

Krista is Your Path to AI Success

The transformative power of Generative AI is undeniable, but the path to successful enterprise deployment is fraught with hidden complexities. Attempting to assemble the pieces on your own risks spiraling costs, security vulnerabilities, project delays, and stifling the very innovation you seek. Krista eliminates these roadblocks. As a purpose-built platform designed expressly for enterprise-wide AI, Krista empowers businesses to capture this technology's potential without the pitfalls of traditional software development approaches. With Krista, you gain speed, security, and adaptability to evolve as AI innovates. Contact us at krista.ai and discover how Krista can accelerate your AI journey, delivering tangible results and a clear competitive edge.

Sources

1. Generative AI is Only 5% of the Solution, Krista Software
2. Enhancing AI Precision with Retrieval Augmented Generation, Krista Software
3. Rise of the Prompt Engineer, Krista Software
4. ChatGPT — Release Notes, OpenAI
5. Comparing Large Language Models for Your Enterprise: A Comprehensive Guide, Krista Software
6. Privacy in the age of generative AI, Stack Overflow
7. ChatGPT costs $700,000 to run daily, OpenAI may go bankrupt in 2024- report, Technext
8. Is Hosting Your Own LLM Cheaper than OpenAI? Hint: It Could Be, Artificial Intelligence in Plain English


Krista™