CHOOSE THE RIGHT LLM: WHY AI FLEXIBILITY MATTERS FOR MAINFRAME TRANSFORMATION

As generative artificial intelligence (GenAI) takes center stage in enterprise innovation, many leaders are asking how to bring its power into their mainframe environments. Some assume it's as easy as plugging in a general-purpose large language model (LLM) like ChatGPT—but this oversimplified approach often misses the mark. The mainframe isn't the problem; in fact, modern platforms like the recently announced IBM® z17™ are fully capable of supporting AI transformation. What matters is how you integrate AI—and more importantly, which AI model you choose.

Building or fine-tuning LLMs from scratch is expensive, time-consuming, and highly specialized. According to a Forrester report, "The State of GenAI in Financial Services," financial services organizations overwhelmingly depend on technology and services partners to deliver GenAI solutions. To succeed, enterprises need more than off-the-shelf chatbots—they need flexibility to select the right LLM or small language model (SLM) for each task, and the ability to adapt as requirements evolve. That's why a curated LLM library and bring-your-own LLM (BYOLLM) strategy is quickly becoming essential for mainframe transformation.

Yet, as with any transformative technology, there's a critical question to ask: What's the right way to implement AI in mainframe environments?

It's easy to assume that selecting an AI model is a one-and-done decision—just pick one, integrate it, and let it work its magic. But AI isn't one-size-fits-all, and the wrong selection can lead to inefficiencies, compliance issues, vendor lock-in, and even security risks. Instead, organizations need

flexibility—the ability to choose, adapt, and control AI in a way that fits their unique business needs.

This is where BMC AMI Assistant is helping enterprises move beyond rigid AI adoption. By providing a curated LLM model library alongside a BYOLLM option, it allows organizations to tailor their AI strategy to their specific workloads, security requirements, and governance policies. But to understand why this level of AI flexibility is so important, we first need to rethink what AI flexibility really means in the enterprise.

What AI flexibility really means—and why it matters

AI flexibility isn't just about having access to different models—it's about ensuring that AI decisions align with business priorities. It means being able to adjust AI strategies as business needs and requirements change, as well as how technology, security policies, and compliance requirements evolve.

Imagine an enterprise that integrates an AI model into its mainframe operations, only to find out months later that it doesn't meet new regulatory requirements. Or consider a company that needs AI to process mission-critical workloads, but the latency is too high, causing inefficiencies that impact the bottom line.

These aren't hypothetical risks; they are real challenges enterprises face when AI strategies lack flexibility. The key to avoiding these pitfalls is having choice, adaptability, and control:

- **Choice:** The ability to select a variety of AI models that best fits the tasks—whether for code explanation, debugging code, analyzing root causes of system issues, or automating responses.
- **Adaptability:** The ability to switch AI models as business needs, compliance laws, and operational constraints shift.
- **Control:** The ability to determine where AI models are hosted, how they interact with data, and how outputs align with enterprise governance policies.

With these factors in mind, enterprises must carefully decide which AI model to use for each task. And that starts with understanding the differences between LLMs and SLMs.

LLM vs. SLM: Choosing the right AI model for the right task

There's a reason why AI leaders don't rely on just one model—different tasks require different capabilities. The choice between LLMs and SLMs comes down to the trade-off between powerful contextual understanding and lightweight efficiency.

Imagine a global bank that needs AI-powered assistance for its mainframe applications. If it wants broad contextual reasoning, such as explaining legacy COBOL code to a new developer, it would likely turn to an LLM—a model trained on vast datasets and capable of understanding relationships across a variety of sources.

But if that same bank needs a highly specific, tightly controlled AI model that generates responses based solely on proprietary company data, it may instead use an SLM—a small, purpose-built AI model designed for speed, security, and precision.

When to Use an LLM

LLMs are powerful tools for scenarios that demand broad, generalized intelligence. They shine when AI needs to:

- Interpret and explain complex relationships across vast datasets.
- Analyze large amounts of unstructured information to identify trends and patterns.
- Support multiple use cases across different business functions, such as customer service automation or IT troubleshooting.

For example, in mainframe environments, an LLM can analyze codebases, detect inefficiencies, and provide AI-driven recommendations to remediate mainframe system issues. But while LLMs are incredibly versatile, they are not always the best fit for high-security, compliance-driven workloads.

When to use an SLM

SLMs, on the other hand, excel when organizations require highly-specific AI models tailored to precise use cases. Unlike LLMs, which are designed for broad generalization, SLMs focus on narrow domains with strict control over data and outcomes. They are best suited for scenarios where:

- Organizations require AI models that are narrowly focused on highly specific use cases, ensuring precise outputs aligned with their unique business requirements.
- Low-latency processing is required, such as in high-speed transaction environments.
- AI outputs need to be tightly governed and based strictly on internal, proprietary data sources.
- Efficiency and cost management are top priorities. SLMs require a lower hardware footprint and reduced computational resources, making them ideal for such environments.

For example, a healthcare organization handling sensitive patient data would likely prefer an SLM for AI-driven documentation rather than exposing confidential information to a broad LLM.

Recent findings from the BMC Mainframe Survey indicate that 64 percent of enterprises identified compliance and security as their top mainframe priority. This highlights the need for SLMs when deploying AI in highly regulated environments.

Understanding when to deploy an LLM versus an SLM is just the first step—enterprises also need the flexibility to choose where their AI models come from. That's where BMC AMI Assistant's curated LLM Library and BYOLLM approach comes in.

AI Flexibility with LLM Library + BYOLM: The Power of Choice

Selecting the right AI model is just one part of the equation. The next question is: What corpus of text was used for training the AI model and what bias could that training contain?

Enterprises benefit from fine-tuned AI models that are ready to use out of the box. Others need to train AI models on their own proprietary data to maintain control over security, compliance, and governance.

BMC AMI Assistant offers both options:

- **Curated LLM Library:** A collection of AI models that are tested and evaluated to work best with BMC AMI Assistant, allowing teams to deploy AI without the burden of building models from

scratch. A core capability is the ease of deployment of LLMs from the AI management console of BMC AMI Platform, ensuring seamless integration and operational efficiency.
- **Bring-Your-Own LLM (BYOLLM):** The flexibility to integrate any AI model best suited to an organization's needs, policies, regulations, and use cases, ensuring full control over security, data privacy, and AI training methods.

For many organizations, the best approach is a hybrid one—leveraging curated LLMs for quick AI adoption while integrating BYOLLM to align to their corporate AI policies. This hybrid approach ensures organizations can adapt AI strategies to their specific use cases, enterprise policies, and security requirements. With the freedom to choose the right model for the right task, teams gain greater control, better alignment with organizational AI policies, and the ability to optimize AI for mainframe transformation.

Adaptability: Keeping pace with change

Adaptability means more than switching between models—it's about aligning AI strategies with the constant evolution of business needs, security demands, and compliance standards. In mainframe environments where workloads are mission-critical, adaptability ensures AI can keep up without introducing risk. As environments change, so must the AI models that support them.

That's why flexibility must include the ability to swap models, retrain where needed, and adopt newer or more specialized LLMs and SLMs over time. With an adaptable architecture, organizations can adjust their AI strategies without rebuilding their systems or compromising performance. BMC AMI Assistant supports this model agility—ensuring enterprises stay resilient no matter how their policies, requirements, or use cases shift.

Is an LLM future-proof?

The AI model chosen today may not meet the business, compliance, and security challenges of tomorrow. This is where a curated LLM Library becomes invaluable. With BMC AMI Assistant, organizations can rapidly take advantage of the latest breakthroughs in LLM technology, ensuring they are always leveraging the most advanced and capable models available. At the same time, they retain the flexibility to pivot—adopting new LLMs when their business needs change, rather than being locked into a single, static AI model.

This ability to dynamically adjust AI strategies ensures that enterprises remain agile, compliant, and ahead of the curve in an era where business needs change and technology shifts at an unprecedented pace.

Final Thoughts: The future of AI for mainframe transformation

AI is actively shaping how organizations manage, transform, and optimize their mainframe environments. However, success in AI adoption isn't just about implementation; it's about ensuring AI remains flexible enough to evolve with the business.

BMC AMI Assistant provides the adaptability and flexibility needed to navigate an ever-changing landscape. With the ability to choose between curated LLMs and BYOLLM, organizations gain the strategic advantage of selecting the right AI model for their needs—today and in the future.

As AI continues to advance, the enterprises that embrace flexibility will be the ones best positioned

for long-term success. The question is no longer whether to use AI in mainframe transformation—but whether the AI strategy in place is built to last.

To learn more about the new capabilities of BMC AMI Assistant—including the curated LLM Library and BYOLLM—read "Transforming the Mainframe's Future with AI-Powered Intelligence," by BMC Vice President of Product Management and Design Matt Whitbourne. You can also discover how the entire BMC AMI portfolio can accelerate your mainframe transformation by visiting the BMC AMI webpage.