REQUIREMENTS FOR BUILDING AN ENTERPRISE GENERATIVE AI STRATEGY

While ChatGPT and GPT-3.5 ushered in a wave of innovations last year, the team behind BMC Helix had already been hard at work for the past few years exploring ways to adapt generative artificial intelligence (AI) technology to enhance enterprise service management applications, improve natural language conversations, emulate human language in chatbots, contextualize knowledge search, and make enterprise service management recommendations for case resolution. The team built several large language model (LLM) prototypes designed to be interoperable with the entire BMC Helix for ServiceOps platform. Our approach to generative AI was purposeful, focusing on the needs of our enterprise customers, and then delivering new use cases that would leverage the technology to resolve problems faster and with greater accuracy.

The result is BMC HelixGPT, a pre-trained generative AI LLM service that integrates into BMC Helix applications, learning from your enterprise's knowledge (including user profiles and permission models) to deliver a tunable, prompt-driven conversational user experience. As 2023 nears the end, we have not only built a scalable generative AI foundation with BMC HelixGPT, but we have also released five new HelixGPT-Powered capabilities:

1. BMC HelixGPT-Powered Helix Virtual Agent
2. BMC HelixGPT-Powered conversations in BMC Helix Digital Workplace
3. BMC HelixGPT-Powered live chat summarization
4. BMC Helix GPT-Powered resolution insights
5. BMC HelixGPT GenAI LLMOps stack, a proprietary generative AI app builder tool (for advanced users)

This blog will be the first in a series that describes our journey in building BMC HelixGPT from end to end and shares key best practices for building a generative AI application with a powerful foundational LLM platform to power it all. If you are building generative AI apps or models, this blog series is for you. We will divide our journey into three parts:

Part 1 (this blog) will focus on unraveling the needs and expectations of a generative AI solution.

Part 2 will outline the components of the BMC HelixGPT platform reference architecture such as those from LangChain, which provide a framework to interact with LLMs, external data sources, prompts, and user interfaces.

Part 3 will show how BMC Helix for ServiceOps leverages BMC HelixGPT to power new enterprise and operations management use cases with leverages BMC HelixGPT's LLMOps capabilities to power enterprise generative AI.

Getting started with generative AI

ChatGPT demonstrated to the world the possibilities of generative AI. What impressed people the most was its ability to quickly provide answers on a vast array of topics in clear, understandable language. Enterprises soon demanded a more tailored approach to the technology, with answers that would be more specific to their internal knowledge and data versus the "world knowledge" that early models were being trained on. To articulate the strategic generative AI direction for BMC Helix, we adopted a systematic three-step process that is universally applicable for enterprises considering generative AI product use cases:

1. Prioritize use cases based on business priorities.
2. Build proofs of concept and get early customer feedback.
3. Understand customer expectations.

Prioritize your enterprise generative AI use cases

One of the initial steps an enterprise must take is to prioritize use cases that align with business goals and priorities based on data availability, customer impact, team skills, and business considerations. In the enterprise service management space, we started with three key use cases that were most impactful for our customers:

- Virtual agent and knowledge search
- Resolution insights
- Summarization

Build proof of concepts and get user feedback early

Once use cases are identified, enterprises need to build proofs of concept to validate the concepts of generative AI. We built customer proofs of concept based on customer data for each of our top three use cases to get feedback through a design partnership program with our customers. While one team was using retrieval augmented generation (RAG)-based approaches and showcasing this to our customers with real customer queries, another team built fine-tuned models for multiple use cases. Early prototypes in resolution insights, generative search, and chatbots were highly impressive and provided us with the learning opportunity to understand and appreciate the

unbelievable power and limitations of LLMs.

Understand enterprise needs and expectations for generative AI

After talking to multiple customers, their expectations of a generative AI solution became clear.

Specificity, accuracy, and trust

Enterprises want answers specific to their data, not generic answers yielded from broad world knowledge. Take, for example, a question regarding resolving a VPN issue, such as "how to fix a VPN connection issue?" Generative AI should generate an answer based on the enterprise VPN articles inside that enterprise and not a generic, plausible-looking answer generated by broad models. They want factual and truthful answers without any hallucinations that are commonplace. Enterprises also want the ability to verify answers with citations or references to build trust in how generative AI sources its answers.

Data security, privacy, and access control

Enterprises have variable user access controls about who can access different levels of data, so answers from generative AI solutions also need to adhere to those same access control policies. For example, a manager and an employee should get different HR answers to the same question because a manager has access to a larger set of documents. A few of the enterprises were also concerned about preserving the privacy of their data and ensuring that it would not be used to train a public model.

Real-time data ingest

Since enterprise data is constantly changing in real time, answers must be also based on the most up-to-date, available knowledge inside the company.

Avoid vendor lock-in of generative AI modeling

Finally, we heard that many enterprises wanted the flexibility to choose their own commercial models like Azure, OpenAI, or open source.

Our early prototypes and high-level requirements collectively shaped the foundational thinking behind what enterprise customers expect from a generative AI solution. In the next blog, I will explain how we addressed customer expectations in our BMC HelixGPT generative AI reference architecture. Stay tuned. In the meantime, you can learn more about BMC Helix GPT here.