WHITE PAPER

Building a Medtech GenAI Platform

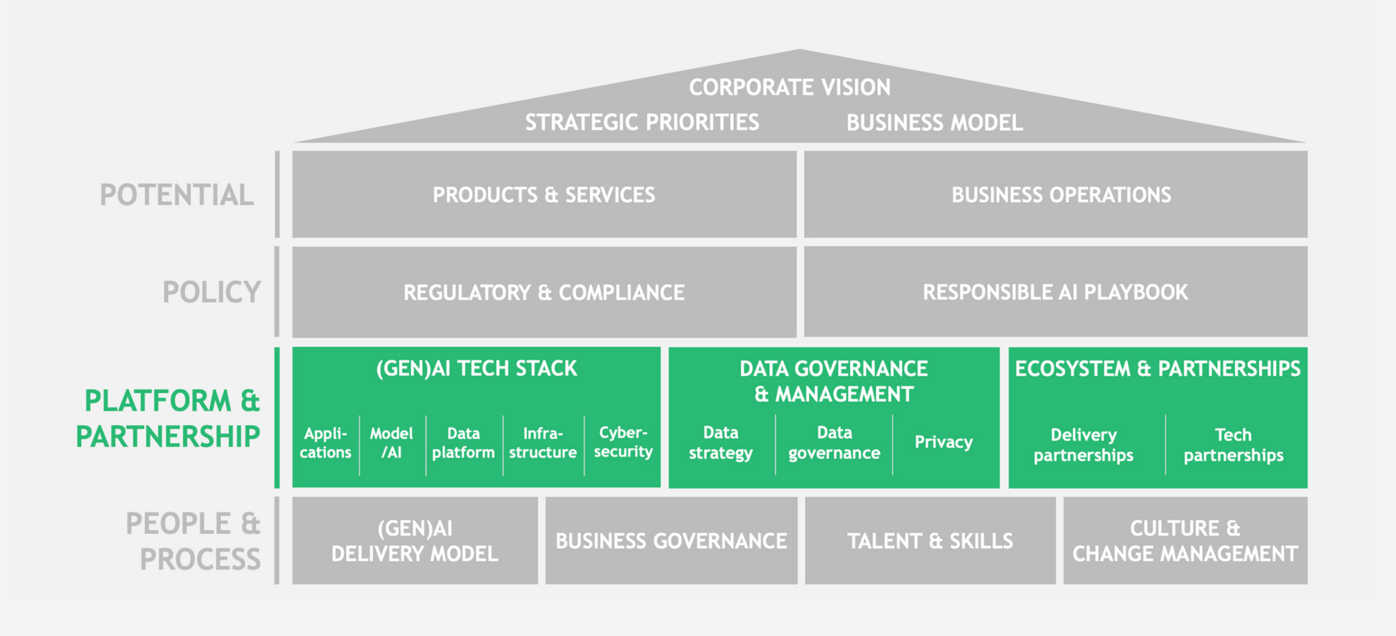
September 26, 2023

By Michael Farrugia, Stuart John, Stefan Leve, Shekhar Marathe, Krishna Srikumar

Building a Medtech GenAI Platform

For all their scientific know-how and feats of engineering with devices and equipment, medtech companies rarely garner praise as front-runners in IT. The emergence of Artificial Intelligence and Machine Learning (AI/ML) has brought new software focus to the industry, with close to 20% of FDA product registrations for AI/ML approved in 2022 alone. GenAI builds on this groundswell, adding a broad range of exciting ways to interact with and augment data, code, images, video, and any other digitized format. This innovative technology can unleash value for all stakeholders (see *Medtech's Gen AI Opportunity*) while raising the stakes for medtech players to get in the game (see *Medtech Companies Must Move Faster on Generative AI*). This article will delve into the key elements of a medtech GenAI platform, exploring the layers of the tech stack as well as the partnership considerations for companies looking forward to fielding GenAI solutions.

This Article's Focus – Platform & Partnership

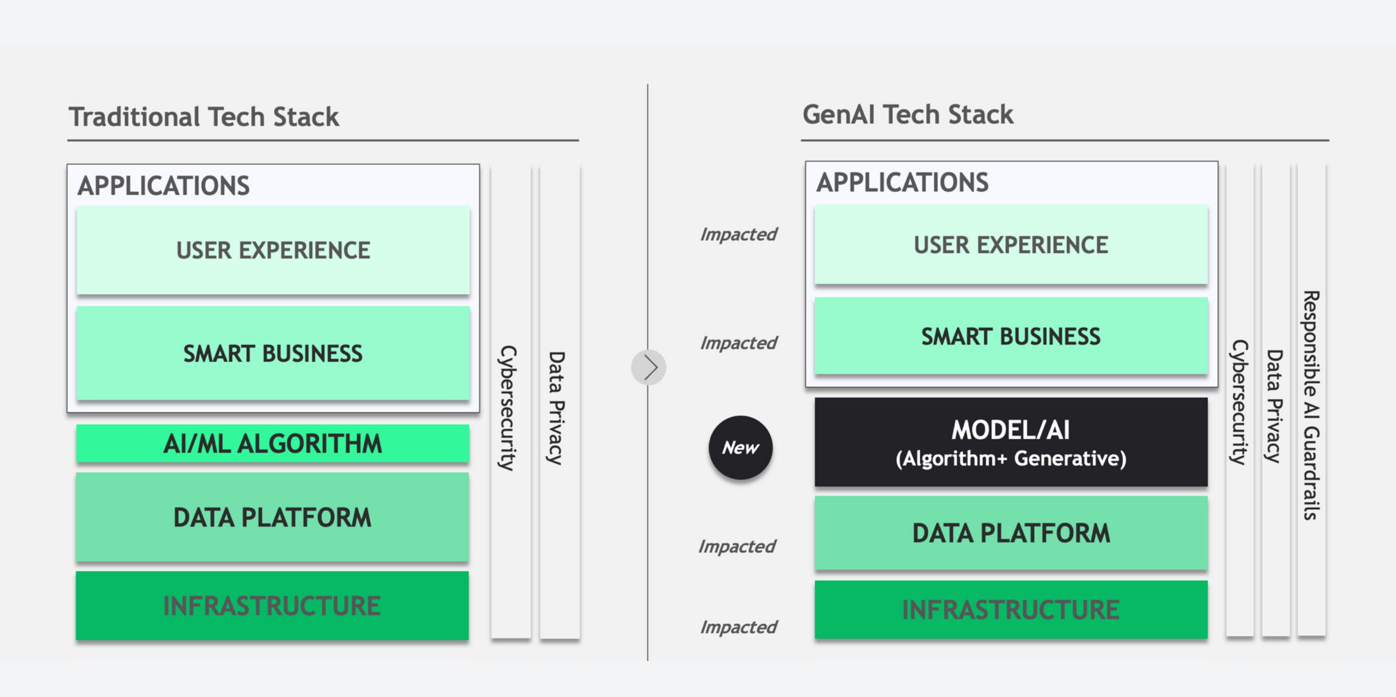


Source: BCG

Layers of the Medtech GenAI Stack

A traditional technology stack for a commercial enterprise includes layers for User Experience, Smart Business services, Data, and Cloud/Infrastructure, each performing a set of crucial functions to deliver common IT solutions. Existing AI/ML algorithms incorporate advanced analytics and reporting or predictive applications, but a GenAI-enabled enterprise requires an additional layer that provides generative capabilities—raising the ante on new functional, technical, and deployment considerations. Medtech companies have to select the right GenAI models for desired modalities (for example, text or image generation, audio text-to-speech or speech-to-text, virtual assistants and chatbots, language translation or software code generation) and their public, managed, or private deployment strategies. When GenAI is added, traditional layers must accommodate the new technology with guardrails that regulate how the model interacts with users and other parts of the tech stack in order to limit unintended or unexpected consequences. Validation and compliance are indelible features of any GenAI solution for medtech—and cybersecurity will be paramount.

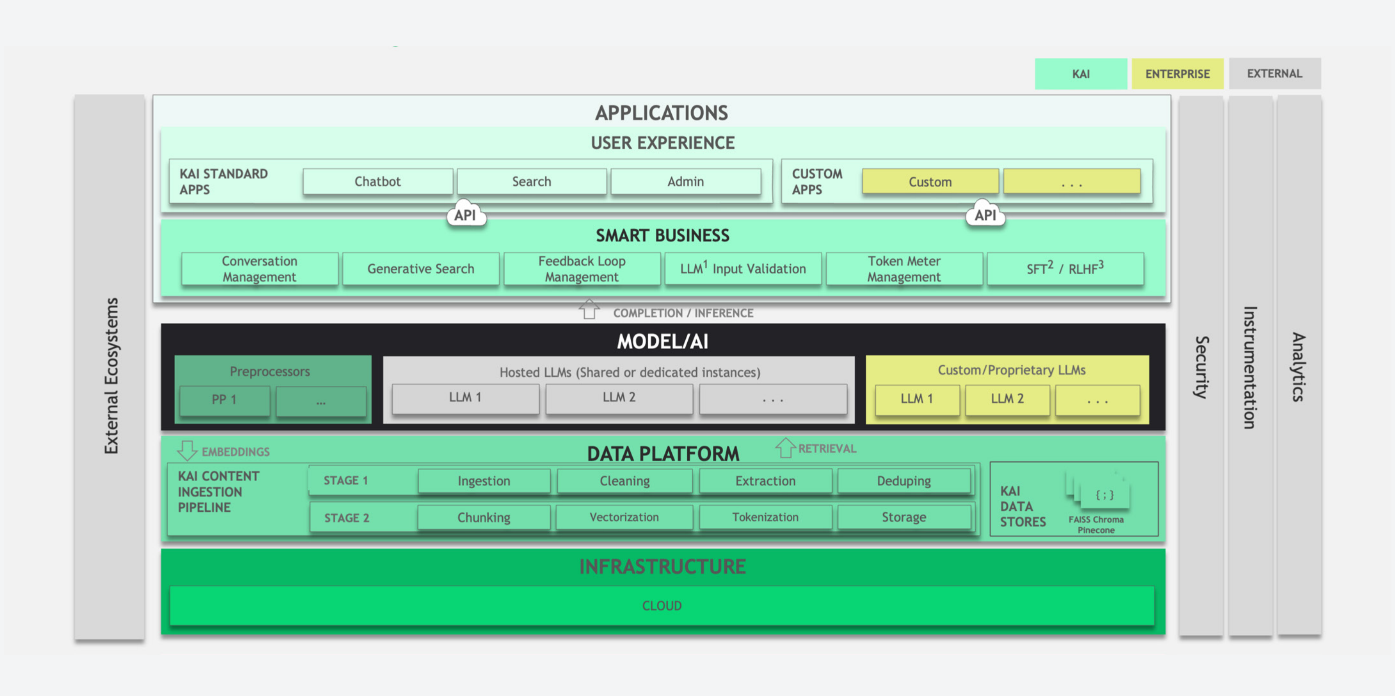
Exhibit 1 - GenAI Introduces a New "Model" Layer to the Technology Stack



Source: BCG
Copyright 2023 by Boston Consulting Group. All rights reserved.

To describe the function of each layer, we use the example of Knowledge AI (KAI), an application built by BCG and deployed at scale for a Customer Support function. KAI is a GenAI-powered application that assists agents in responding to customer requests quickly through multiple channels by generating the precise information and instructions for problem resolution, based on training with an expansive set of data and documents.

Exhibit 2 – Knowledge.AI Tech Stack



Source: BCG
Copyright 2023 by Boston Consulting Group. All rights reserved.

KAI is fully integrated into the enterprise IT security context to ensure data privacy and protection. The application acquires documents and contextual data from a variety of internal and third-party sources in both structured and unstructured formats, processing them for consumption by generative models. The model incorporates instrumentation and usage analytics for observability and system performance analytics. The key components of KAI and other GenAI systems include the following:

- **User Experience (UX) Layer.** Much of the buzz around GenAI stems from chat interfaces with underlying large language models (LLMs), such as ChatGPT, which enable users to ask questions using natural language instructions, or “prompts.” Medtech companies can use this conversational capability to create tailored and engaging user experiences. For example, KAI can accept a detailed description of a problem in a customer’s own terms and then retrieve and synthesize the most relevant and specific answers. The generative model can also format the answer to fit a customer’s specific need, often combining text with images and videos to illustrate the recommended solution. The UX allows a Customer Service rep to provide feedback about the accuracy and relevance of the solution so that KAI can learn and improve based on this reinforcement learning from human feedback technique.

- **Smart Business Layer.** This layer provides orchestration and automation frameworks and “low code” tools that enable solution staff developers and vendors to design and deploy client applications. This layer typically makes use of open source frameworks such as Langchain to combine multiple LLMs, which in turn generate and validate outputs. Other common features include the ability to author, test, and refine prompts and manage context, templates, or examples. For KAI, BCG developed knowledge-retrieval services that manage and recall successfully resolved chat sessions to train both agents and the system on the optimal resolution path for customer concerns, enabling tremendous improvement in first-time issue resolution and cost savings.

- **Model/AI Layer.** This layer can build on existing AI/ML algorithms, offering new capabilities with general purpose Foundation Models that are immediately useful across a broad set of functional uses, as well as Specialized Models with specific explainability requirements and guardrails for sensitive clinical applications. Companies will find it simplest to experiment with off-the-shelf LLMs such as OpenAI's GPT-4 to achieve fast implementation of text-based use cases (or applications of models such as Github Copilot/Codex for code and test-case generation). With additional experience and confidence, medtech companies may "fine-tune" specialized LLMs with a smaller, structured data set that distills proprietary data insights into a variety of clinical applications including personalized treatment plans with AI-assisted health coaching.

  Medtech companies may combine single-purpose AI/ML algorithms, such as detection, diagnostic, or risk-prediction models, with generative models that interpret the output to write clinical notes and offer clinical recommendations for providers and patients to take action. KAI's chart-generation feature uses an LLM to generate both a textual description of the data and code for extraction and visualization of data tables into charts and graphs, which could be deployed within a medtech product for clearer explanation and education around the data.

- **Data Platform Layer.** A GenAI platform moves beyond traditional AI/ML by supporting multiple modalities such as visual, audio, and textual data—often in combination—to enable a more diverse range of applications. However, this shift brings additional challenges for data quality, privacy, and storage—ensuring data provenance, classification, lineage, and compliance become essential to ensure the reliability and accuracy of GenAI outputs. Similarly, bias and discrimination can arise from compromised training data (such as data not intended for this use or copyright protected) that skews model output and justifiably attracts regulatory scrutiny.

  Existing medtech AI/ML data layers typically incorporate three functions: A Data Acquisition and Ingestion function allows a model to acquire millions of data points across different health events generated by devices in a defined user population. The model must have a Data Storage and Processing capability that records and interprets different health events for various patient types in device readings, rendering structured output for further analysis. Also, the model relies on Data Inference to instruct the provider, patient, or device to take specific action (for example, changing device settings or seeking specific medical attention). GenAI can enrich existing medtech AI data sets used for training and testing by generating synthetic data to augment sparse real-world information, mitigate bias, preserve and enhance data privacy, and simulate future health scenarios. Generative models can also create new metadata based on inputs, for example, by interpreting images or video frames and providing annotations for further analysis and assessment by clinicians. A fundamental processing step employed in this layer by KAI is taking unstructured from data sources such as reports, user guides, and technical manuals and extracting it into structured data tables as contextual input to models where visualization of the provided data is one of the generated outputs.

- **Infrastructure and Cloud.** Model deployment and infrastructure play a crucial role in determining a GenAI platform's efficiency and performance. While it is technically possible for a company to train a private model on a large-scale data set, this may be cost prohibitive until synthetic data generation becomes more efficient. The typical configuration in the short term—which BCG employed for KAI—is to access a large pretrained language model through a managed service provider, such as Microsoft Azure, with a pay-for-consumption agreement. This has some advantages over connecting directly to a proprietary Application Programming Interface (API) such as XML and trying to enforce customized data protection and privacy policies. Existing and emerging Responsible AI tools that monitor and measure fairness and explainability are pre-wired to interface with leading LLMs, enabling developers to deploy fine-tuning data sets and fully functional applications more rapidly. For now, companies may be obliged to support a multi-cloud architecture, as specific models may only be available from a single cloud platform provider.

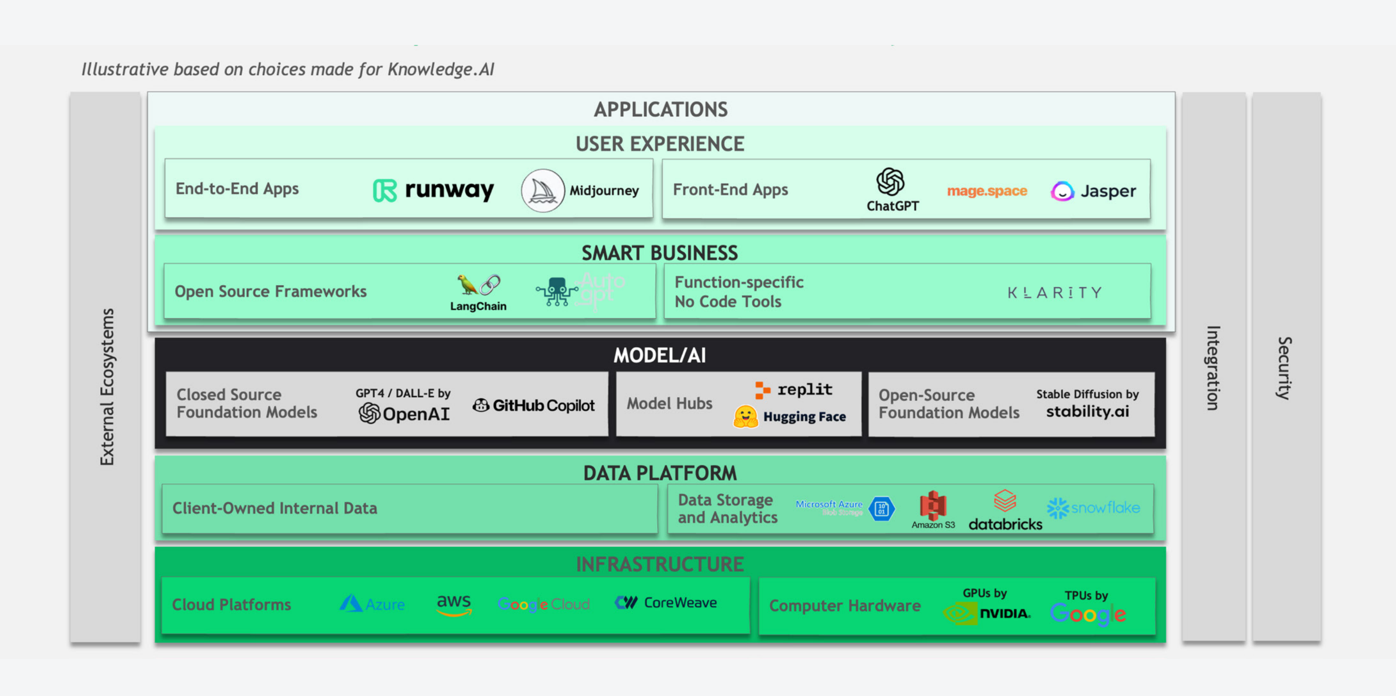
Selecting and Managing GenAI Partners

Given that many emerging GenAI models and tools are open source, medtech executives may not view developing partnerships as a core element of their GenAI. But tech partnerships can help a company's GenAI effort get to scale with existing pre-trained Foundation Models, hit the market sooner by leveraging a partner's unique skills and capabilities, and reduce the up-front capital requirements and risks entailed in developing a suite of GenAI offerings.

Partnering also provides inherent future-proofing in a swirling ecosystem, where partner selections today—particularly long-term arrangements with one-stop shops—could become a millstone in the years to come. In the short term, companies should expect to have multiple provider ecosystem partners, extending current managed platform contracts to include GenAI capabilities, but also incorporating new partners that offer differentiated model and tool choices.

A partners' ability to uphold a Responsible AI framework, including an organization's policy guardrails and risk-monitoring strategy, should be a crucial determinant of their suitability. A partner that is truly on board will enable a medtech company to develop GenAI use cases that feature compliant and secure privacy-by-design controls through the software development life cycle.

Exhibit 3 – Partnerships to be Evaluated for Each Layer



Source: BCG
Copyright 2023 by Boston Consulting Group. All rights reserved.

In addition to the functional and technical choices for generative model selection, the key determinants of an effective GenAI partnership include the enabling services and support required to manage the rapid evolution of a GenAI platform:

- **GenAI** Use Case Requirements. Functional or operational improvements that are not specific to the medtech industry can be handled by a pre-trained LLM model that is fine-tuned to recognize and manipulate a general range of data inputs within the context of a company's policies, guidelines, and enterprise-specific knowledge to generate outputs more closely aligned with their unique ways of working while minimizing the risk of "capability overhang" or unintended uses and consequences of general models. Conversely, clinical functions may require a more specialized domain-specific model embedded with privacy measures and tight controls on source data to guard against bias and hallucination. Nascent "model hub" providers are emerging to offer a library of models for a variety of use case purposes, supported by training and a community for collaborative model development.

- **Cloud and Security Preferences.** The emerging thicket of regulatory requirements across countries and regions concerning data hosting imposes conflicting requirements and constraints. Consequently, medtech companies may need to deploy multiple data storage sites and draw on different cloud partners to manage GenAI services. Though redundant, such mixed-model scenarios offer advantages beyond compliance—especially in the realm of cybersecurity. With data storage isolated and GenAI use cases optimized for local/regional deployment, medtech companies will require support for input validation in the form of tailored detection logic, data queries, and malware decomposition analysis to expose direct attacks, malicious traffic, and anomalous behavior. GenAI models can be tuned to support rapid, or automated, technical review to incident response actions and output validation. Output validation ensures the output of the model is in line with enterprise and responsible AI policies while minimizing the risk of hallucination.

- **Cost.** The cost of GenAI use cases depends on the commercial structure established by model vendors, cloud providers, and other computational resources. Per-use fees represent the most common payment structures, though some vendors, such as Midjourney and ChatGPT Plus, offer flat monthly rates for dedicated capacity or subscription offers. Consumption of GenAI resources is typically measured in "tokens" of input and output. In addition to these charges, medtech companies must parse the attractiveness of volume discounts and incentives for longer-term contracts against the risk of getting locked into an unsatisfactory arrangement.

Platform Evolution: Getting to GenAI

Building a GenAI stack clearly requires incremental additions to existing platform architecture as well as a new set of choices for models, services, and tools. To gain prompt access to this innovative technology, medtech players can benefit from partnerships with clear leaders in the field such as OpenAI, Microsoft, Google, and other crucial enablers.

GenAI models themselves are only useful when embedded into an application that is suitable for purpose, informed by user needs, established with clear business priorities, aligned with the organization's principles and policies, and supported by the layers of the tech stack. The first stage of understanding GenAI involves ad hoc experimentation and developing proofs of concept with LLMs to understand the possibilities and assess the business value of the technology.

As a shared understanding develops around initial pilots, the next stage of the GenAI journey will revolutionize specific functions, workflows, and use cases. Medtech executives need to be thinking now about how these expected developments will impact and interact with the policies and guidelines enshrined in their company's Responsible GenAI framework. Embracing GenAI and harnessing its vast potential while hewing close to regulatory requirements, core values, and social responsibility will enable medtech organizations to emerge as leaders in this promising and fresh new frontier.

Acknowledgments

The authors would like to thank Peter Lawyer and the following for their contributions to this article: Ian Chen, Peter Lawyer, Dan Martines, Tom Retelewski, Stephen Spencer, and Gunnar Trommer.