



HPE AI Services —Generative AI Implementation

Develop your own generative AI project and run it to address your organization needs



What it looks like when industry practice is applied

Once you have identified the use case of reference driving the application success and related data that can come from your organization or public sources, the next natural step is to experiment with the implementation, gain insights on the value generated, and evolve and bring it to the core of your organization's business.

Depending on the needs and available resources, you have a wide set of options to strategically select with HPE AI Services—Generative AI Implementation Service. Should model generic abilities be enough for your case, HPE AI experts implement the needed deployment model and optimize for inference, resulting in a ready-to-use solution to consume and integrate with the existing processes or deploy as a stand-alone new application (see Figure 1).

When domain verticalization is needed, or complex contexts are provided as input to the system, effort is required with a minor impact on resources and a high focus on quality data and business value. It is accomplished by applying prompt engineering on top of inference tasks.

Moving forward with operationalization, a continuous feedback loop involving human evaluation is extremely important for keeping the model performance at the initial level. For this reason, we apply Advanced Optimization Techniques (AOT) developed by HPE AI expertise gained in years of experience working on production-ready AI solutions.

Why generative AI? Why now?

Generative AI and the adoption of Large Language Models (LLM) represent one of the big artificial intelligence (AI) advancement waves.

The ability to let machine systems generate controlled outputs that are natively human readable tremendously simplifies their use, accelerating market acceptance and experimentation enthusiasm. The key factors building the current success are:

- Novel model architectures such as transformers, leveraging the existing deep learning principles and available technology resources with innovative structures to address existing areas such as natural language processing
- Open-source project proliferation, defining new job roles working on optimized frameworks to improve and optimize capability and on model specialization with advanced optimization techniques and scalable fine-tuning approaches
- Large consumer availability, making cutting-edge discoveries available to anyone, including profiles with high creativity and less technical background, accelerating the exploration of use cases

Finally, if there are specialized features that need to extend the model capabilities, it is required to build fine-tuned versions from the model of choice involving a moderate use of computational resources. It is accomplished by applying additional training either by updating the entire model or a portion of it, depending on how disruptive the desired additional specialized feature is for the model.

We are open to organizations' creativity, and should a completely new model architecture be required or the need to use specific training sets arises with generative AI and LLM design and build, we revisit the model architecture and retrain on targeted and accordingly prepared data.

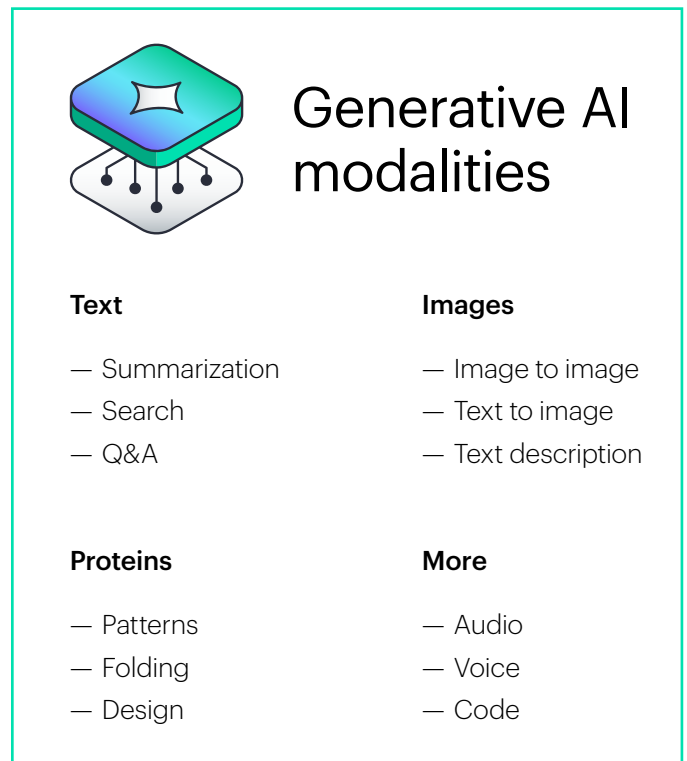


Figure 1. Generative AI modalities



What to expect

Our AI experts bring data science, machine learning (ML) engineering, and ML Ops expertise to assess the business expectations, align to technology requirements and performance metrics, and prepare the required data in a secure and proper format to be consumed by the models (see Figure 2).

Data is the foundation of any AI project's success. Dedicated pipelines help ensure the smooth transition between source, platforms, and output destination while loaders are used to handle human-readable representations, which are translated into machine-readable formats keeping semantics and related context.

Technology provides the needed resources to prepare and run the deployed solution, optimized infrastructures, and cloud-native software platforms that help team collaboration and agile development lifecycle.

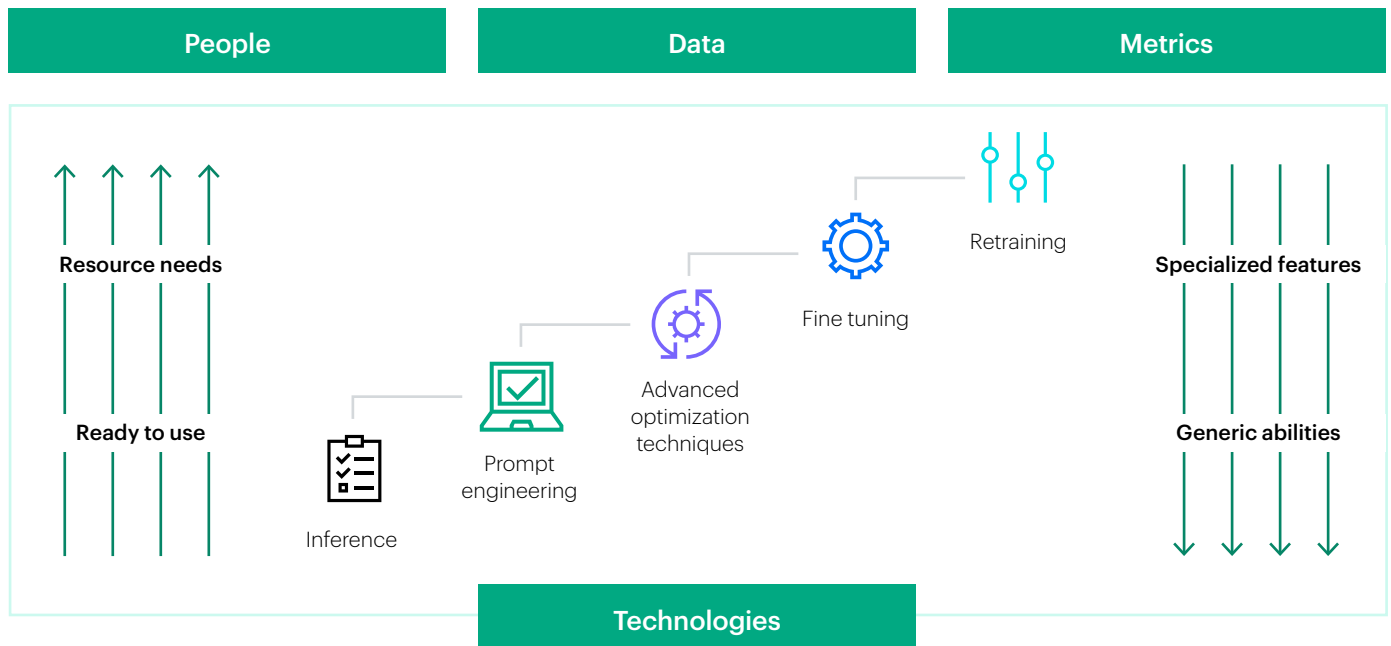


Figure 2. Key engagement elements

What data, analytics, and AI advisory and professional can provide

- **HPE AI Services**—Expertise to explore, experiment, and evolve your AI adoption journey
- **HPE Private Cloud AI**—Turnkey, scalable, and AI-optimized private cloud designed to accelerate AI project deployment while helping ensure data remains under enterprise control
- **HPE GreenLake**—Edge-to-cloud platform delivering AI, ML and data analytics, high-performance computing (HPC) as-a-service on-premises, in the cloud, or in a colocation facility near you
- **HPE compute and storage**—Modern, HPC solutions for GPU, data-intensive workloads, edge/IoT analytics, and secure data management solutions
- **Hewlett Packard Labs**—Innovations such as Trustworthy AI and cookbooks for DL workloads based on extensive benchmarking
- **Partner ecosystem**—Technology and cloud partners offering data and analytics solutions

Ecosystem

Open-source projects and partner solutions are introduced when needed to accomplish the business objective. HPE AI experts integrate and enhance the ecosystem offerings, building and deploying a unique solution functional to your desired purposes.

We offer the option to accelerate outcomes leveraging NVIDIA software with the selection of NIM based use cases deployment for optimized inferencing setup and unmatched serving performance, and with NeMo Framework widening the scope to data preparation, models customization, guardrails.

Learn more at

- [HPE.com/ai-services](https://hpe.com/ai-services)
- [HPE.com/GreenLake](https://hpe.com/GreenLake)

Visit [HPE.com](https://hpe.com)

Chat now

© Copyright 2025 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

a00134579ENW

HEWLETT PACKARD ENTERPRISE

hpe.com

