Ready, set, deploy!
GenAI implementation options

One of the most important decisions you'll have to make in your generative AI (GenAI) journey is which implementation pattern to use. As you move up the Enterprise AI Continuum toward GenAI, your data has a greater chance of getting into the public domain, which means you need stronger security protocols to protect it. As you prepare to pull the trigger on your GenAI strategy and use cases, how to address increasing security demands is a key factor in determining which GenAI implementation pattern is best suited to your environment and your goals.

Security, of course, is not the only issue to consider as you plot your GenAI course forward. Governance, cost, scalability, ethics and risk tolerance are important factors as well.

Our team has assessed a number of different GenAI implementation patterns against these criteria and others over the past year-plus. We've settled on the four we think are the most promising – and the most pragmatic – for the majority of enterprises. Each one has its pros and cons and all of them require you to take a business-centric view of your goals for deploying GenAI.

1 2 3 4

1 COTS Point Solution

Description:
Commercial off-the-shelf solution that has GenAI built in

Provider examples:
Amelia for conversational AI (i.e. chatbot)
Otter.ai for meeting summaries

Best used for:
Narrow and well-defined use cases
Augmenting existing software capabilities

Benefits:
Least expensive (usually) and time-consuming to implement
Not much implementation risk

Downside:
Likely requires separate solutions for different use cases, which can add up to be a management or IT burden and chip away at cost benefits

Cost: $-$$$

Bottom line:
Let your software vendors add GenAI as a feature and source it from them.

2 Public Access LLM

Description: Direct access to public version of LLM (e.g., ChatGPT) with a custom-built "filtering application" that screens prompts and responses for content violations

Provider examples: Bard (Google), ChatGPT (OpenAI), LLaMA (Meta)

Best used for:
General information summarization where no sensitive data is included or required
Code generation
Synthetic data creation
Marketing/sales planning and content

Benefits:
Filtering application is the only custom-built software you'll need
Low implementation risk

Downside: Limited ability to include proprietary data securely

Cost: $-$$

Bottom line:
As long as you mitigate the potential data security risk and/or you don't need your own data embedded in the model, this is a relatively fast, easy and cost-effective option.

3 Private Instance LLM

What:
Bring LLM foundational model (e.g., OpenAI's GPT-3.5) into your own environment
Vector database that stores proprietary data
Custom-built orchestrator to embed proprietary data into user's prompts

Provider examples:
Microsoft Azure OpenAI, Amazon Bedrock, Google's Gemini

Best used for: GenAI custom solutions that include secure and sensitive data – e.g., knowledge base queries, financial insights generation or custom content generation

Benefits: Potentially the least expensive option for including rich proprietary data

Downside:
Need an existing relationship with the LLM provider
Requires custom-built orchestrator and processes to maintain/update LLM

Cost: $$-$$$

Bottom line:
If you need to include your own data and you already have a relationship with an LLM provider, this is potentially your path of least resistance.

4 Custom-trained LLM

What:
Fine-tune public LLM or custom-train an open-source LLM
Vector database that stores proprietary data
Custom-built orchestrator to embed proprietary data into user's prompts

Provider examples: N/A

Best used for:
GenAI solutions in highly regulated industries
Use cases where accuracy and repeatability are paramount (e.g., employee, patient or client safety)

Benefits: Maximal customization and compliance

Downside:
Potentially costly and time-consuming to train, manage and maintain

COST: $$$-$$$$

Bottom line:
If compliance and/or data protection are your top priorities, then a custom-trained LLM will likely be worth the cost and effort.

What's your pattern for success?

NTT DATA chose **Pattern #3** – the Private Instance LLM – for our own internal GenAI architecture. As part of this effort, we've built accelerators that clients can use to deploy similar elements including a reference architecture, sample orchestrator and tools for creating embedding jobs.

As we see the private instance LLM grow in popularity – and with the NTT DATA accelerator offerings now available – it's worth exploring to see if it fits your needs. It's definitely not for everyone as it can be more expensive to manage and maintain. But with NTT DATA's experience building out the architecture, its prototype accelerator to boost content security, and our partnership with many of the leading LLM providers, it could turn out to be the right GenAI engine for your organization's success.

But, for clients that are interested in either **Pattern #2** alone or in conjunction with one of the other patterns, we've also built an accelerator for a custom content filter that is a mechanism for implementing a client's GenAI Acceptable Use Policy. It functions as a content filter to protect your IP, PII and any other proprietary data that you don't want to get into the public domain. It can also protect employees that are using AI from things like abusive language or inappropriate responses.

At the end of the day, there is no one right answer for every enterprise. The key is to make enough progress on your GenAI strategy to get a sense for the highest value use cases that would indicate the best implementation pattern (or set of patterns) for your enterprise.

Contact our AI Advantage team
to identify the approach that best suits your organization and its goals.