softserve

GENERATIVE AI / THE RACE IS ON

FOUR ADOPTION PATTERNS FOR YOUR ENTERPRISE

THE FUTURE IS NOW

Pundits have called generative AI technology a once-in-a-lifetime moment — a transformative technology to rival the commercialization of the Internet. The hype is deafening, even confusing.

Industries across the business spectrum have entered the race to understand technology and determine where it can provide the most value. And while the generative AI journey is not a sprint, executives must approach generative AI with a heightened sense of urgency.

While many have spent countless words talking about what Generative AI is and why it matters, few explain where to start and how to make Generative AI real in the enterprise.

SoftServe's adoption patterns aim to provide you — business leaders — with just that: the voice of clarity to navigate the fog of emerging technology with a chance to avoid pitfalls and pass through the trough of disillusionment, reaching the plateau of Gen AI productivity faster and more successfully than your competition.

GENERATIVE AI ADOPTION PATTERNS

The multitude of options and specialized terminology in the Gen AI field can be perplexing. To help you puzzle out and choose the right approach for your Gen AI endeavor, we describe four adoption patterns here and depict them in the diagram below:

PAY AND USE

Rapidly adopt third-party application, but with limited customization and control

INTEGRATE YOUR APPS

Seamlessly integrate third-party Gen AI model into your application(s)

ENRICH WITH YOUR DATA

Customize third-party Gen AI model with your data and integrate it into your apps

TRAIN ON YOUR DATA

Develop custom Gen AI model trained on your data & integrate into your apps for full IP ownership, control, and customization

PAY AND USE

ChatGPT by OpenAI is undoubtedly the most well-known 3rd party Gen AI service nowadays, having reached 100 million monthly active users in January, just two months after its launch. Today, the field of Gen AI SaaS applications is booming, offering users a range from universal chatbots like ChatGPT and Claude.ai by Anthropic, to more specialized tools such as GitHub Copilot for drafting code. The following table shows key aspects of the Pay and Use Generative AI Tools approach. This table will help you compare the SoftServe approach with other options and better understand if it is right for your business.

USE CASE EXAMPLES	The most suitable use cases: • Code drafting for Software Engineering; • Content drafting for Marketing and Sales. Somewhat suitable use cases: • Extracting information from documents for various corporate roles. The least suitable use cases: • online customer support.
LIMITATIONS	• Accuracy, truthfulness, and inherent bias are common issues of Gen AI models • App specific limitations, e.g. ChatGPT cannot access data or current events after September 2021; • The primary limiting factor is that in this business pattern 3rd party Gen AI apps are not integrated into the corporate ecosystem, which restricts their value to individual usage. They provide limited or even transient competitive advantage, as such tools are widely available.
PROPRIETARY DATA	Proprietary data might be used in manual prompts or when uploading documents to SaaS apps, such as Claude.ai.
SECURITY/PRIVACY	User input may be stored for the purposes of future model training. As for copyright on the generated content, laws can vary depending on the jurisdiction and the specifics of the use case, and policies may change over time. For example, under current U.S. law, AI-generated content is not considered to be an original creation of the human mind and is therefore not protected by copyright law.
TECHNOLOGY	Foundational or specialized models available for end-users as web, mobile apps, or plugins.
COST	Typically fixed-fee subscriptions of $10 to $30 per user per month, some products have usage-based pricing.
IMPLEMENTATION TEAM	This adoption pattern does not require a technical team.

OUR RECOMMENDATION

Evaluate 3rd party Gen AI apps, adopt them for corporate roles where they display value, and monitor their further development. The field is emerging, and for example, it is anticipated that GPT-5 will be 10 to 20 times larger than the current GPT-4, which could enable new opportunities for your business.

In the meantime, let’s explore more value that lies in integration with your apps and corporate ecosystem.

INTEGRATE YOUR APPS

The good news for businesses is that popular LLM foundational models are available through APIs, which can be integrated with existing applications or used for building custom software. In the table below, we will explore key aspects of the “Integrate Your Apps” pattern:

USE CASE EXAMPLES	The most suitable use cases: • Document processing and information extraction for various corporate functions, such as procurement, risk, compliance, and legal; • Sentiment analysis; Somewhat suitable use cases: • Online customer support, virtual agents; The least suitable use cases: • Corporate knowledge base assistant.
LIMITATIONS	See limitations in “Pay and Use”. In addition there are technical limitations related to APIs: • The size of the context window (the data that the model can "see" or consider) is limited and varies from service to service. For instance, as of now, GPT-4 supports up to 32K tokens. As a rough estimate, you can think of 1 token as roughly equivalent to 1 word. This corresponds to about 100 pages. • While this pattern allows for integration with an application's data context, the primary limitation is that it does not leverage corporate data stores such as wikis, document management, data lakes, or data warehouses. The next two patterns aim to solve this limitation.
PROPRIETARY DATA	Proprietary data from corporate apps are passed to a 3rd party Generative AI service in the form of prompts.
SECURITY/PRIVACY	3rd party terms can vary but most Cloud vendors do not transfer customer’s data to other customers and do not use it for improving their models.
TECHNOLOGY	Azure OpenAI Service (GPT-4, gpt-3.5-turbo), Google Cloud (PaLM 2), Amazon Bedrock (Titan, Claude 2, etc.), Nvidia (NeMo). Most services provide REST API and client libraries for Python, Java, Java Script and other popular languages.
COST	The monthly cost of usage 3rd party Gen AI service heavily depends on the amount of passed and generated information (tokens), for example a virtual 24/7 agent answering around 10,000 questions per day would cost from $210 to $6,000 monthly (based on GPT-3.5 and GPT-4 Turbo current pricing). At the same time, there are options for using open-source models that can noticeably decrease the total cost of infrastructure for high-load applications. The cost of integration with corporate apps will depend on the number of those apps and their complexity; thus, it is not included in the calculation.
IMPLEMENTATION TEAM	Solutions Architect, Software Engineers, DevOps Engineer

OUR RECOMMENDATION

Carefully assess your existing applications and identify where Gen AI integration can provide the most value. Consider the technical limitations, such as the context window size and the cost based on the leveraging selected model and API usage. Depending on the option, the cost difference can vary by up to 30 times. When the need arises in the corporate knowledge base, then the next pattern is worth looking into.

ENRICH WITH YOUR DATA

Looking to use existing LLM solutions with your own data? No worries, it doesn't necessarily require training your own LLM. LLMs introduce a cost-effective approach to knowledge utilization called embeddings. Through AI-powered indexing and search, GenAI systems can access information from corporate systems such as portals, wikis, data warehouses, data lakes, and other sources of structured or unstructured data. The integration of LLMs with proprietary data opens the door to leveraging the distinct knowledge amassed by an organization — a resource uniquely their own.

USE CASE EXAMPLES	The best use cases for this pattern are the ones that rely on corporate data and knowledge, e.g., Customer Service Automation, Marketing Content Generation, Auto-reporting, Enterprise Search, Job Specific Assistants/Copilots, and Decision Support systems.
LIMITATIONS	• LLMs enable enhanced content indexing using AI-powered embeddings (indexes). However, it also introduces the need for an additional data infrastructure for processing and storing the embeddings; • Fitting corporate knowledge into the model's context could be challenging. However, adding an extra processing logic may help extract concise summaries from the most relevant documents.
PROPRIETARY DATA	Additional proprietary data from corporate data sources is passed to an embedding model for content indexing. Self-hosted open-source embedding models can be used to secure sensitive data.
SECURITY/PRIVACY	In addition to the embedding model, corporate data is also passed to a 3rd party LLM, potentially exposing sensitive data to users via an LLM response. Proper guardrails and response filtering must be implemented to prevent sensitive data exposure.
TECHNOLOGY	Azure OpenAI, Amazon Bedrock, Google Cloud Embedding API, Cohere, and Nvidia NeMo are the most popular 3rd party services for embedding models. HuggingFace Hub provides a wide range of open-source embedding models. Milvus, FAISS, Chroma, and Weaviate are the most popular open-source embedding databases, and Pinecone is a good 3rd-party alternative.
COST	Like in the abovementioned pattern, “Integrate Your Apps,” the monthly cost of usage depends on the amount of passed and generated information. However, in this case, indexing data through embeddings also incurs additional costs. Although the cost of embeddings seems low ($0.0001 per 1K tokens for GPT), and indexing 1 million 30-page documents would cost just about $1600, indexing 1 petabyte of text may easily exceed $10M when using a 3rd-party service. Thus, using open-source models for embedding can drastically decrease the cost when supporting massive data.
IMPLEMENTATION TEAM	Solution and Data Architects, Software, Data, and ML Engineers, DevOps Engineer.

OUR RECOMMENDATION

Leverage corporate data as a powerful differentiator. Consider utilizing open-source models to mitigate expenses, especially when dealing with large volumes of data. Implement robust guardrails and response filtering to prevent potential exposure of sensitive data. In case your system relies on specific domain expertise, the next approach — Train on your Data — can be a better option.

TRAIN ON YOUR DATA

The most value can be extracted from LLMs when it’s trained on the specific data relevant to the business and subject domain. This pattern is more expensive to implement but provides more accurate responses for specific domains or skills. The good news is that the initial model training can be eliminated by using pre-trained Foundational Models (FMs) and fine-tuning them on the proprietary data. Some third-party party LLMs allow fine-tuning the model on the customer's data, while others do not. In the latter case, the model can be fine-tuned on the customer's data using self-hosted open-source FMs.

USE CASE EXAMPLES

The use cases for specialized skills or specific domain (Legal, Medical, Finance, etc.) are the best candidates.

LIMITATIONS

- Training and fine-tuning LLMs requires a significant amount of training data and extensive compute resources (including GPUs/TPUs).
- Data quality is a critical factor for the success of this pattern – smaller but higher quality data brings better results. Additional data curation and governance processes may be required.
- Many 3rd-party LLMs are limited in training data customization (e.g., GPT-3.5 and GPT-4 do not support fine-tuning on the customer's data, only older versions).
- LLMs trained on a specific domain may not be able to generalize to other domains.

PROPRIETARY DATA

In addition to user and corporate search data, fine-tuning a 3rd party LLM requires passing a significant amount of data relevant to the business and subject domain for knowledge extraction.

SECURITY/PRIVACY

3rd party terms may vary but most Cloud vendors do not share customer’s data and fined-tuned models between tenants.

TECHNOLOGY

Google Vertex AI, Amazon SageMaker, Azure Machine Learning, HuggingFace Hub, and NVIDIA NeMo are the most popular services for training and fine-tuning LLMs. HuggingFace Hub provides a wide range of open-source Foundational Models.

COST

The cost of fine-tuning varies heavily depending on the approach: managed service vs. open source. For example, fine-tuning foundational models with biomedical data (~50B tokens) using the OpenAI Davinci model costs $1,500,000, compared to $43,550 when using the open-source MPT-30B. For comparison with fine-tuning, here are some examples of training foundational models from scratch:

Scenario 1 (commodity model)

- Architecture: GPT3-style
- Number of parameters: 70B
- Hardware: A cluster of 256 x Nvidia A100-40GB
- Training dataset size: 1.4T tokens
- GPU Hours: 1,084,723 hours
- Approx. Cost: $2,500,000

Scenario 2 (state-of-the-art model)

- Architecture: Llama-2
- Number of parameters: 70B
- Hardware: GCP, a2-ultragpu-1g, NVIDIA A100 80GB
- Training dataset size: 2T tokens
- GPU Hours: 1,720,320 hours
- Approx. Cost: $8,719,958

IMPLEMENTATION TEAM

Solution and Data Architects, Data Scientist, Software, Data, and ML Engineers, DevOps Engineer

OUR RECOMMENDATION

The “Train on Your Data”approach is more complex and expensive but offers the greatest benefits. Leverage pre-trained foundational models and fine-tune them on proprietary data to reduce initial model training costs. Pay attention to data quality and consider additional curation and governance processes, as smaller but higher quality data may yield better results, and be mindful of the significant costs and resources required for training from scratch.

GET STARTED ON YOUR GENERATIVE AI JOURNEY

LET SOFTSERVE ACCELERATE YOUR GENERATIVE AI JOURNEY WITH 3 OFFERINGS

Don't be left behind in the race to harness the disruptive innovation of Generative AI. SoftServe's adoption patterns are designed to **help you navigate the complexities of this emerging technology**. We will help you select an appropriate adoption pattern or its combinations for building tailored solutions to meet your enterprise's unique needs. Expedite your journey towards a new way of productivity and creativity with SoftServe Generative AI offerings, positioning your business as a leader in the industry:

AI DISCOVERY

Interest to Discovery: Generative AI Ecosystem and Implications for My Business

- Use Cases & Business Impact Priorities
- Data Quality and Availability
- Technology Trade-offs and Architecture
- Technical Feasibility with POC

AI LAUNCHPAD

Launchpad to Innovation: Evidence-Based Exploration and Deployment

- Generative AI Lab
- AI Launchpad Program for rapid experimentation
- Value Stream Mapping & Use Cases
- POC/POV Pipeline

AI ADOPTION

Insight to Impact: Rapid Scaling and Adoption in My Organization

- Generative AI Adoption Roadmap
- Technology Strategy
- Data Strategy
- Change Management and AI Governance

- Generative AI Solution Development
- Generative AI in Product and Engineering Teams

It's urgent and the time to start is now. But you know that Generative AI is not a deployment sprint. With SoftServe's experts to address your unique enterprise needs and guide you through possible adoption options, you will move past the hype and harness its benefits for your enterprise.

Learn more about SoftServe's Generative AI Lab, POVs, offerings, and partners.
Visit us at our Generative AI website page.

YOUR PARTNER FOR THE FUTURE

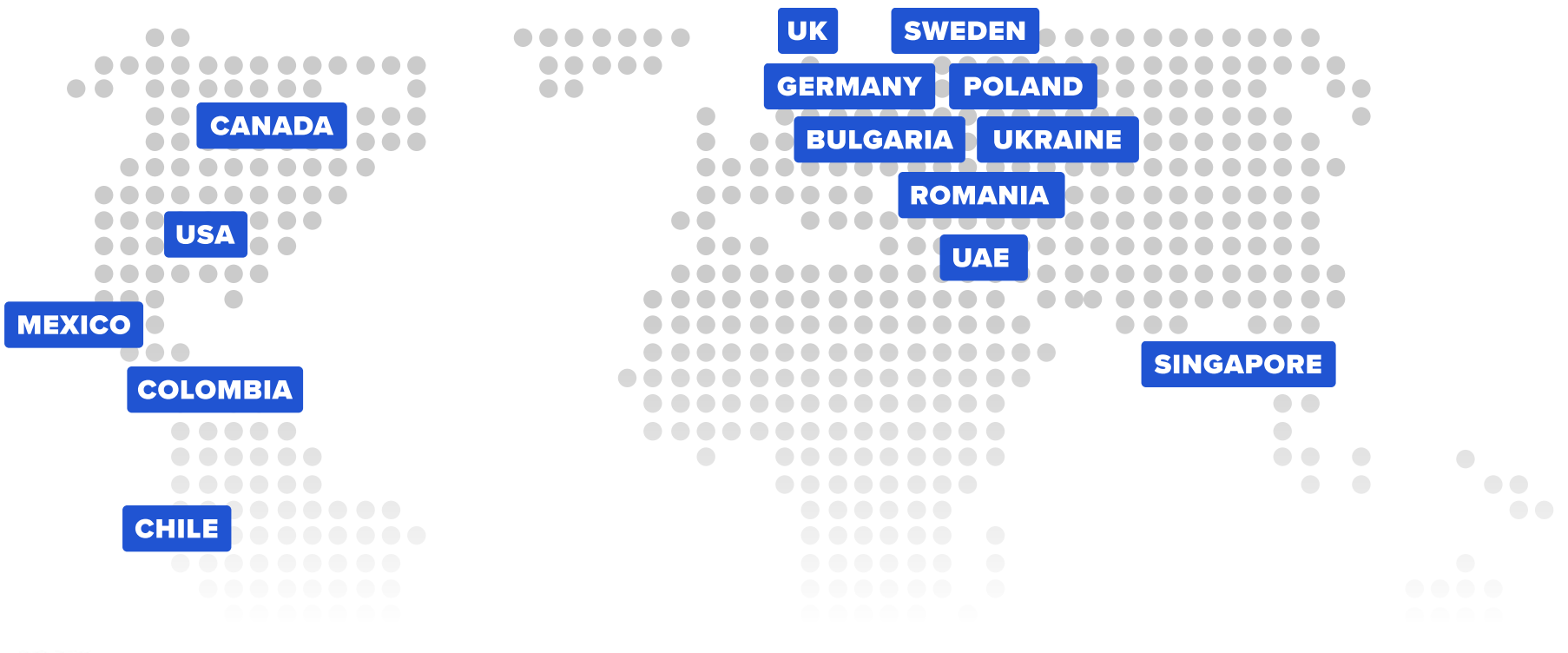
DIGITAL ADVISORS, SOFTWARE INNOVATORS, AND EXPERIENCE BUILDERS

30 Years of Experience

98% Client Retention

14 Countries of Operations

58 Offices Worldwide



GOOGLE CLOUD EXPERTISE

As a Google Cloud Partner and the winner of the 2020 Google Cloud Partner of the Year Award for Machine Learning, SoftServe is committed to helping customers solve their most pressing business challenges. SoftServe utilizes Google Cloud's AI ecosystem, including tools such as Google Cloud Vertex AI and Generative AI App Builder. Learn more

120+ GCP Enabled Customers

690+ GCP Certified Recources

100+ Data Science Experts

3 Anthos Fellows

AMAZON WEB SERVICES

As an APN Premier Services Partner, SoftServe acts as an exceptional cloud guide, vastly decreasing the time to achieve cloud value. By doing so, SoftServe ensures that your AI initiatives unleash the full potential of AWS Machine Learning services, such as Amazon Bedrock and SageMaker, and that they are deployed in accordance with AWS Well-Architected best practices. Learn more

500+ AWS Certified Resources

100+ AWS Certified Solution Architects

50+ AWS Professional Certifications

11 AWS Competencies

MICROSOFT

Innovate with purpose, rationalize costs, and drive efficiencies with Microsoft Azure's open and flexible cloud computing platform. Leverage Azure OpenAI and Machine Learning services to deliver next-generation AI solutions. As a Gold Microsoft Partner, SoftServe enables your business to build and deploy on your terms — both today and in the future. Learn more

250+ Azure Certified Professionals

19+ Years as a Microsoft Partner

NVIDIA

As an NVIDIA Service Delivery Partner, SoftServe harnesses NVIDIA's cutting-edge technologies, like GPU-accelerated compute infrastructure, to deliver robust AI solutions. Leveraging NVIDIA's NeMo Service, SoftServe streamlines the development of Generative AI products, driving rapid digital transformation. Learn more

softserve

DEEP TECHNOLOGY EXPERTISE IN AI/ML

HUMAN-FIRST APPROACH TO DESIGNING EFFECTIVE EXPERIENCES

INDUSTRY EXPERTISE AND PARTNERSHIPS WITH CLOUD SOLUTIONS PROVIDERS

Iurii Milovanov,
AVP AI/Data Science, imilov@softserveinc.com

Serge Haziyev,
Advanced Technologies CTO, shaziyev@softserveinc.com

Alex Chubay,
Chief Technology Officer, ochubay@softserveinc.com