AIM | RESEARCH

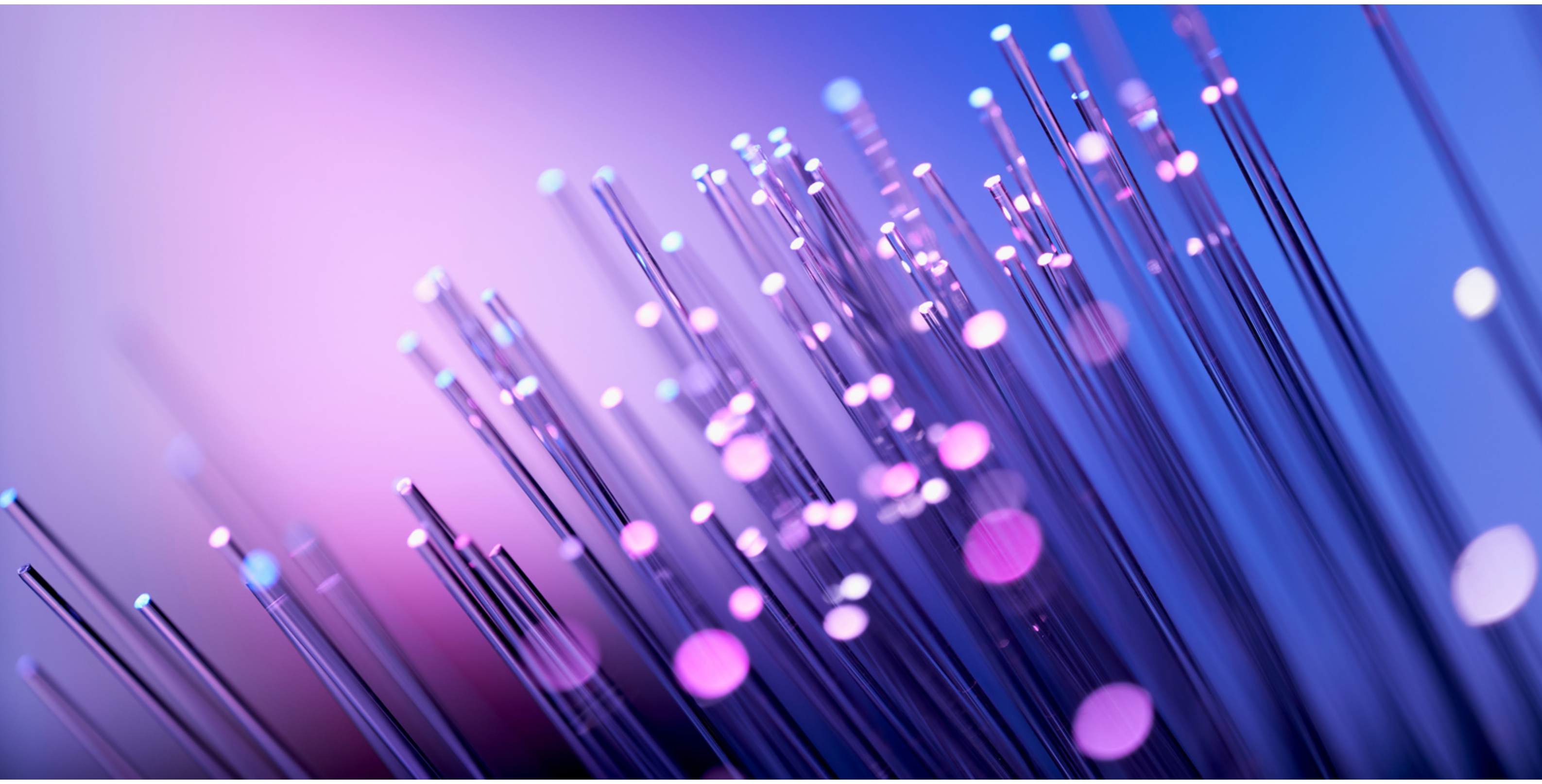
hansa cequity
ENRICHING CUSTOMER EQUITY
ISO/ IEC 27001:2013 CERTIFIED

AIM Research GenAI Insights

LLM Economics - A Guide to Generative AI Implementation Cost

Conversations with industry experts and C-suite executives illuminate the evolving landscape of generative AI, offering critical insights into the complexities and variables that influence implementation costs in today's dynamic economic climate.

by Ayush Jain



September 2023

Table of Content

Foreword

Generative Artificial Intelligence (AI) has gained significant attention for its potential to transform various industries. Some of the ways that an organisation can use generative AI are - Personalising customer experiences, streamlining operations and efficiency, enhancing decision-making, preserving privacy and security, fraud detection and cybersecurity. However, most organisations are encountering challenges when implementing generative AI in their systems. Understanding the costs involved and developing sustainable solutions is crucial for organisations looking to leverage generative AI effectively.

Hansa Cequity has been at the forefront of helping clients implement bleeding edge technology solutions for more than a decade. This Guide to Generative AI Implementation Cost in collaboration with AIM Research aims to provide valuable insights and guidance for organisations looking to leverage generative AI effectively. By combining primary and secondary research methods, we have analysed industry trends, assessed cost components, and explored best practices. We have also focused on marketing-driven examples and case studies to showcase practical applications.

In this Guide, we will provide you with an overview of how to conduct a cost-benefit analysis for generative AI projects. We will cover the following topics:

1. A Case Study Approach to Generative AI in MarTech Lifecycle
2. Cost Analysis
3. How to Reduce Cost
4. Roadmap for Implementation

We hope that this guide will help you to make informed decisions about generative AI implementation and to maximize its value for your organization and more specifically build a roadmap for sustainable implementation of text based LLMs. Generative AI also poses some challenges and risks, such as data quality, ethical issues, legal implications, and social impact.

Therefore, before implementing generative AI solutions based on a cost-benefit analysis, it is also important to conduct the feasibility, viability, and desirability of the project. Overall, the applications of generative AI are vast and varied, and it has the potential to transform many different industries.

IAs the technology continues to advance, it will be interesting to see the new and innovative ways in which it is used in the future. Hansa Cequity along with AIM Research will be keeping a close eye on this fast-evolving space. I am sure you will find this Guide practical and useful.

Neeraj Pratap Sangani
CEO, Hansa Cequity

hansa cequity

ENRICHING CUSTOMER EQUITY

ISO/ IEC 27001:2013 CERTIFIED

Executive Summary

As we find ourselves amidst the 'ChatGPT moment', LLMs stand at the fulcrum of a transformative wave, prompting industry leaders to regard this development as a powerful tool to '**reduce costs and increase profits**'. But, the market doesn't seem to be unfolding as per the hype. A comprehensive understanding of the infrastructure necessary to maximize the potential of this new domain—including insights into the cost-benefit ratio, pertinent use cases, and the motivations driving organizations to adopt such tools—remains elusive.

On top of that, Gartner's recent research also forecasts a significant slowdown in enterprise deployments in the general AI space. As highlighted in the study, it is projected that over the **next two years**, the overwhelming costs will exceed the value that will be generated, culminating in about **50%** of the large enterprises abandoning their large-scale AI model developments by 2028.

To get to the crux of reality, AIM Research hosted a roundtable discussion comprising of several AI leaders from different industries working in this space. Here are some key insights that came to light:

- Identifying the appropriate use case with quantifiable business benefits is critical. It involves understanding the technology's capabilities and aligning them with business objectives.

- Starting with a POC allows businesses to evaluate the potential impacts before scaling up. It is also crucial to be aware of the costs involved in scaling up, including cloud and API usage costs.

- A sensible approach to budgeting would involve allocating more towards improving operational efficiency initially through AI integration to optimize processes, cut costs, and improve service levels, and as the system matures, gradually shift funds towards customer acquisition strategies, utilizing AI to enhance personalization and engagement.

- For AI success, organizations must focus on improving prompt engineering for targeted insights, and excel in data fusion to combine various data sources for more accurate and useful information, promoting collaboration and integration within the organization.

- The future of AI seems to be leaning towards agent technology, where multiple AI agents work together to achieve specific tasks, instead of a single AI entity handling all tasks. These technologies would be industry-specific and would collaborate similarly to a human mind, although achieving this level of integration and function is still a far-off goal.

- Organizations are evaluating both API and open-source options for AI integration, weighing factors like speed to market, customization, and regulatory requirements. While APIs might be favored for pilot projects due to their quick deployment, open-source might be the choice for full-fledged production, offering better audit facilities and customization options.

Thus, a report like this could serve as a vital tool in this process, helping stakeholders to assess the potential costs and benefits associated with different implementation strategies, whether it be through API or open-source pathways. It could clarify the complexities of both direct and indirect costs, facilitating smarter decisions that consider factors such as quick deployment and customization options.

Ultimately, such a report could guide organizations in choosing the most suitable and cost-effective solutions for AI integration.

Introduction

While the allure of Generative AI in enterprise solutions is undeniable, there exists a cloud of uncertainty surrounding its actual costs of implementation. Many enterprises struggle with the less concrete parts of using AI, like getting to know the complex technology, the unpredictable nature of generative models, and issues related to data privacy and control. However, the main worry is about the financial aspect.

The direct and indirect costs associated with integrating AI into production systems are ambiguous, often leading to misconceptions. **For businesses, especially SMEs, deciphering these costs is crucial**, from initial deployment to the long-term aspects of maintenance, updates, data management, and security. The rapid evolution of AI technologies further compounds this challenge, as models need frequent monitoring, updating, and retraining to stay effective.

This ambiguity calls for comprehensive research that can demystify the various components influencing the cost of implementing Generative AI. **By breaking down the myriad elements, from external APIs to self-hosting open source models on Cloud, a clearer picture can emerge,** dispelling myths and giving organizations a more grounded understanding. Such research would offer a reality check against the myriad numbers often cited, and guiding enterprises in their AI endeavors with more precision and confidence.

The cost of implementing Generative AI can vary immensely, often dictated by the specific industry use case, the modality of the model, and a plethora of other factors. For instance, an AI model designed for a healthcare application—where precision is paramount and errors can have grave implications—might demand more rigorous training, higher-quality data, and specialized expertise than, say, a model used for generating text in a blogging platform. Therefore, it becomes important to first define the scope of any research aimed at understanding the costs of AI implementation.

> "I believe it's crucial to begin and experiment. Given this is a new space - from a CXO's perspective, the right thing to do would be to focus on internal use cases initially - as they would carry less risk. There are numerous scenarios where this can make a significant difference. Start there to build momentum and gain experience, and then transition to tackling more impactful use cases - including customer facing ones."

Arvind Mathur, Chief Information Officer AMEA at Kellogg's

You can also access our custom-built LLM cost calculator here.

Scope of the Report

The research will delve into the use cases typically observed in MarTech functionalities. Additionally, while there are categories of models that can produce produce text, images, videos, and even voice, this report will limit its focus only on **text-based Large Language Models** (LLMs).

Methodology

The research design for this study employs a case study approach. We will consider four use cases from different industries within the Martech lifecycle and calculate estimated costs under various implementation scenarios. This research will then be validated through secondary studies, consultations with industry experts, and focus groups. The multi-faceted approach bridges the gap between theoretical estimations and the practical realities of developing these models for enterprise use cases.

Additionally, for each use case, we estimate the cost of developing it as a chatbot.

> "From a marketing communication perspective, I expect generative AI implementations to happen sooner. This is because they don't require constant changes. I envision numerous use cases emerging within the next year, with a lot more industries coming up around that as well."

Roshan Thayyil, Head of Loyalty Analytics at Emirates

Research Objectives

The research objectives for this study are as follows:

- Assess the cost implications of implementing generative AI in production systems
- Identify the key cost components, including infrastructure, data acquisition, talent acquisition, training, and maintenance
- Investigate the challenges and roadblocks that consumer companies face when integrating generative AI into existing workflows
- Explore strategies and best practices for building sustainable generative AI solutions while optimizing costs
- Analyze industry trends and patterns in the adoption of generative AI and associated costs

How to read this report

Determining the average cost in Generative AI is not a straightforward task. Therefore, this report is structured to guide you progressively through this nuanced topic in the following way:

Section 1: **Defining the Use Case** - We lay the groundwork with descriptive case studies that shed light on different real-world applications. This is to arrive at a realistic estimation of how many tokens are generated for each of the use case.

Section 2: **Cost Analysis** - Next, we conduct a detailed cost analysis, focusing on both API and Cloud GPU pathways to provide a balanced view. At the end of each analysis, you'll see a visual representation of the up and above cost beyond simple integration.

Section 3: **Strategies to Reduce Cost** - In this section, we explore potential strategies to trim costs effectively without compromising output quality.

Section 4: **Roadmap for Sustainable Implementation** - Finally, we propose a forward-thinking roadmap, sketching a path for sustainable and economically viable generative AI implementations.

We have also developed a cost calculator tool to facilitate an easy estimation of approximate costs for your specific use case. You can access this tool [here].

Chapter 1

Taking a Case Study Approach

Our case studies will cover each stage of the customer lifecycle — acquisition, retention, engagement, and win-back — and focus on four distinct applications of generative AI within the MarTech sector. For each of the case studies, we will estimate the cost of generation using external API and self-hosted models.

MarTech: The Current Hotbed

Coca-Cola’s recent ad, "Masterpiece", was made partly with generative AI, featuring a combination of film, 3D, and Stable Diffusion techniques. While the film was brilliantly executed, it wasn't cheap! It underwent numerous rounds of testing because it needed to have a "pull dimension".

Like the Coca-Cola case, there have been numerous examples within the MarTech sector where efforts are weighed in terms of "effectiveness" (content generation that appeals to consumers) rather than "efficiency". ChatGPT has proven to be a significant influencer in demonstrating how it can impact this function.

A McKinsey report indicated that players that invest in generative AI are seeing a revenue uplift of **3 to 15 percent** and a sales ROI uplift of **10 to 20 percent**.

At the same time, as per AIM Research, if we look at the rate of adoption by function, we see that sectors like Automotive & Manufacturing, Retail & CPG and Pharma & Healthcare are experiencing **20%, 20%, and 17%** adoption of generative AI in marketing respectively. Marketing & Sales form the highest compared to all other functions.

The report will, therefore, utilize use cases from various stages of the MarTech lifecycle to better contextualize the incurred costs and provide an understanding of how the approximate cost is determined.

> "One development we'll likely witness is more rigor towards separation between generative AI for efficiency and generative AI for effectiveness. When considering advertising content or other messaging, the shift toward effectiveness will demand more resources and usher in a greater sense of accountability. In that sense, we will see a move towards more A/B testing in terms of content."

Arvind Balasundaram, Executive Director, Commercial Insights & Analytics at Regeneron Pharmaceuticals

Exhibit 1

Implementing Generative AI Solutions Across the MarTech Lifecycle

	Use Case	Industry Examples	Prompting/Finetuning Technique	Timeframe
Acquisition	Sentiment Analysis for an eCommerce Company	RedCloud's optimized ad spend for FMCG brands	Few-shot learning, in-context learning	6-9 months from concept to deployment
Retention	Tailored Content in the FMCG sector	Amazon's tailored content; Shopify's "Magic" feature	Adaptive sampling, temperature tuning	5-8 months from initial brainstorming to deployment
Engagement	Customer Complaint Redressal in the Automotive Industry	BMW's AI-driven generative design system	In-context learning, continuous learning	6-10 months from ideation to practical application
Win-Back	Churn Analysis & Incentive-Based Win-Back in BFSI	JPMorgan Chase's IndexGPT application	Few-shot learning, temperature tuning	8-12 months from research to on-ground implementation

Comparing External API and Self-hosted Models

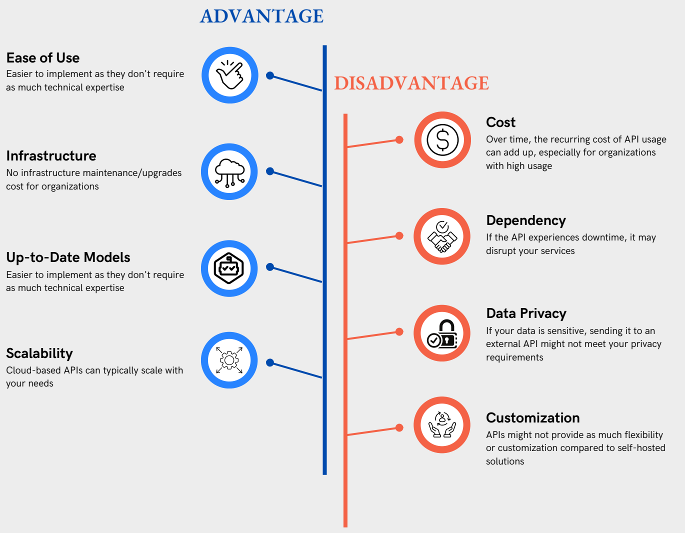
Deploying generative AI into organizations' processes can indeed be done in two main ways: using external APIs or self-hosting large language models (LLMs) on Cloud. Both these methods come with their own advantages and disadvantages, and your choice would largely depend on your organization's specific needs and constraints.

Exhibit 2

Comparing External API and Self-hosted Models

External API

External APIs refers to an external provider that offers pre-built AI models accessible via APIs, allowing developers to leverage AI capabilities without the need to train or host their own models.



ADVANTAGE

Ease of Use
Easier to implement as they don't require as much technical expertise

Infrastructure
No infrastructure maintenance/upgrades cost for organizations

Up-to-Date Models
Easier to implement as they don't require as much technical expertise

Scalability
Cloud-based APIs can typically scale with your needs

DISADVANTAGE

Cost
Over time, the recurring cost of API usage can add up, especially for organizations with high usage

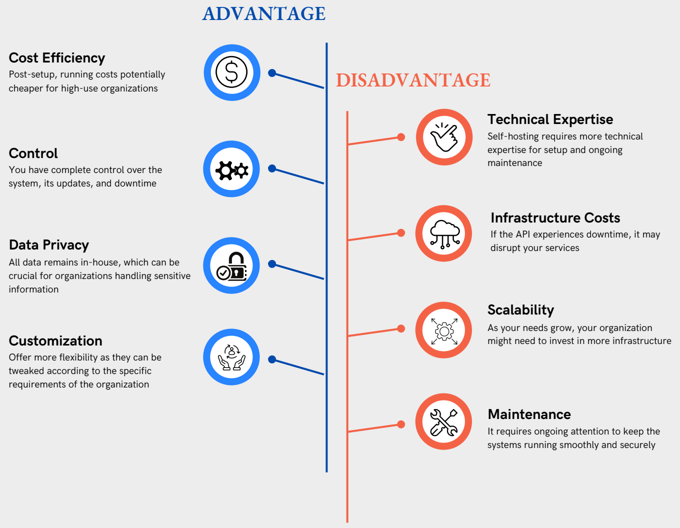
Dependency
If the API experiences downtime, it may disrupt your services

Data Privacy
If your data is sensitive, sending it to an external API might not meet your privacy requirements

Customization
APIs might not provide as much flexibility or customization compared to self-hosted solutions

Self-hosted Models

Self-hosting language models on Cloud is a feasible solution for those who want to have their own deployment without relying on third-party API services. This enables more control, potential cost savings, and possibly better performance.



ADVANTAGE

Cost Efficiency
Post-setup, running costs potentially cheaper for high-use organizations

Control
You have complete control over the system, its updates, and downtime

Data Privacy
All data remains in-house, which can be crucial for organizations handling sensitive information

Customization
Offer more flexibility as they can be tweaked according to the specific requirements of the organization

DISADVANTAGE

Technical Expertise
Self-hosting requires more technical expertise for setup and ongoing maintenance

Infrastructure Costs
If the API experiences downtime, it may disrupt your services

Scalability
As your needs grow, your organization might need to invest in more infrastructure

Maintenance
It requires ongoing attention to keep the systems running smoothly and securely