AIM RESEARCH

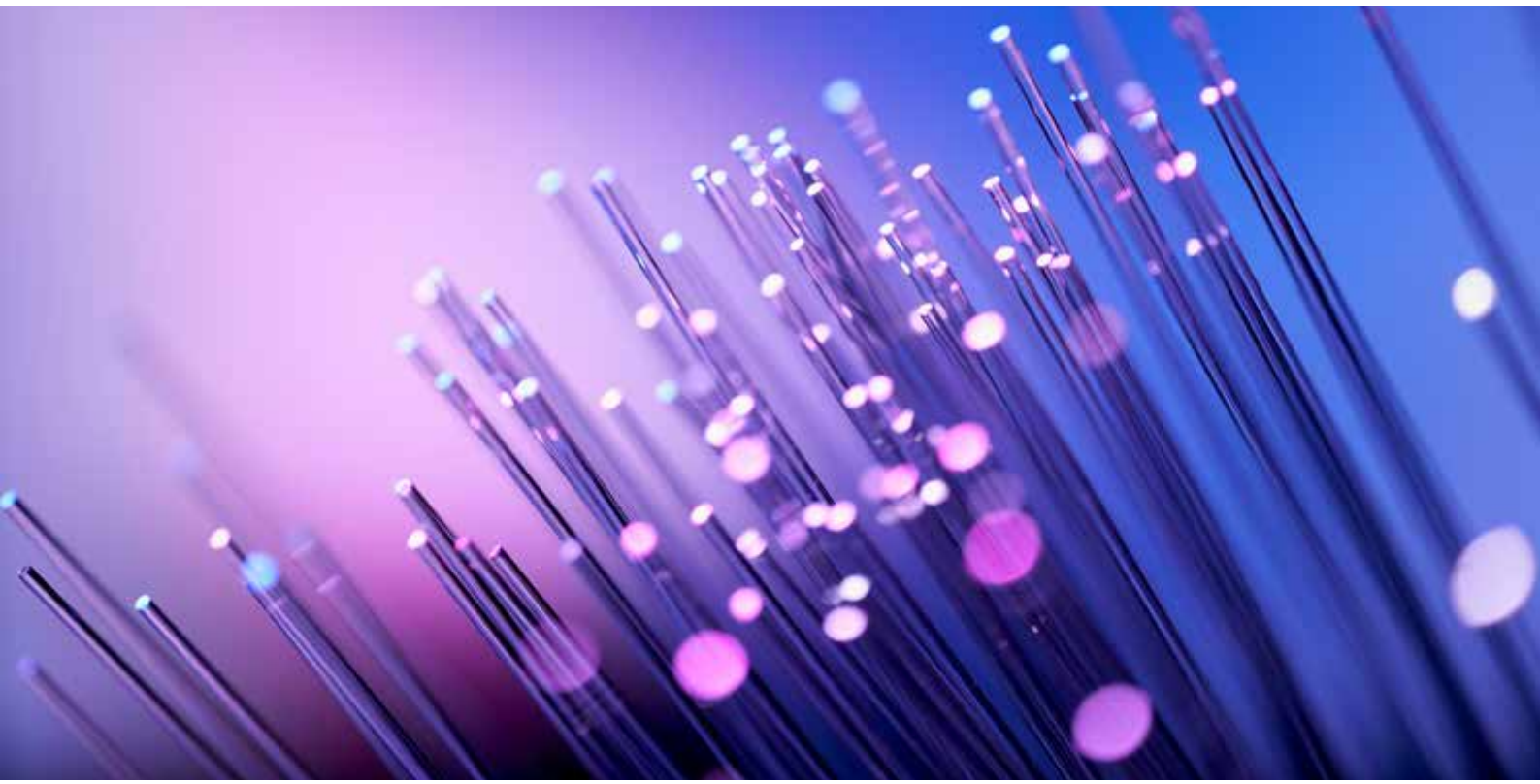
hansa cequity
ENRICHING CUSTOMER EQUITY
ISO/IEC 27001:2013 CERTIFIED

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

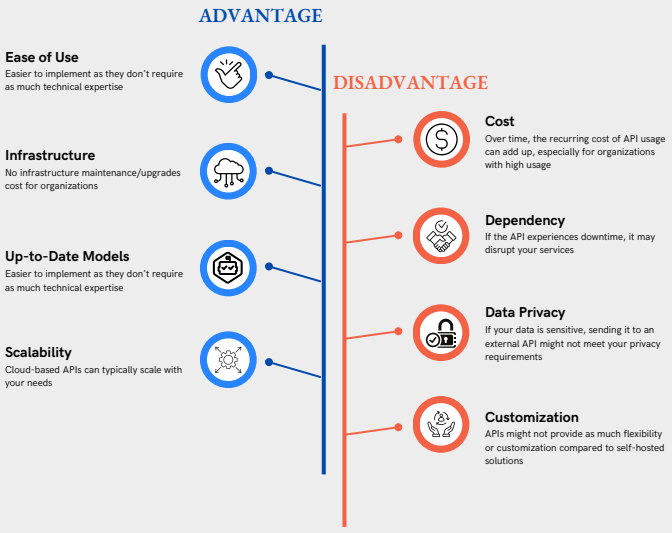
Comparing External API and Self-hosted Models

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

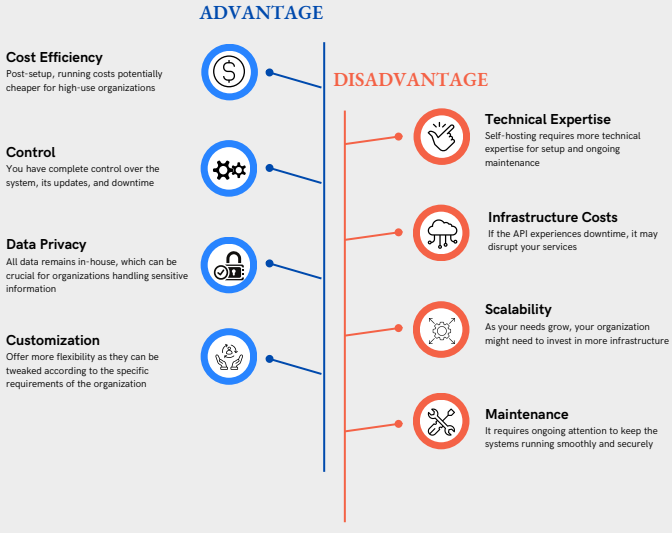
External API

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.



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.



"The first and foremost question is identifying the right use case that offers a quantified business benefit. Additionally, it's essential to consider if you truly seek 100% accuracy. If you aim for perfect accuracy, this field might not be suitable for you at the moment.

Cost is another significant factor. Whether you're working at the proof-of-concept (POC) level, using cloud services, or utilizing APIs, costs are typically calculated per token. Hence, we can make informed estimates about potential expenses for a given amount of data.

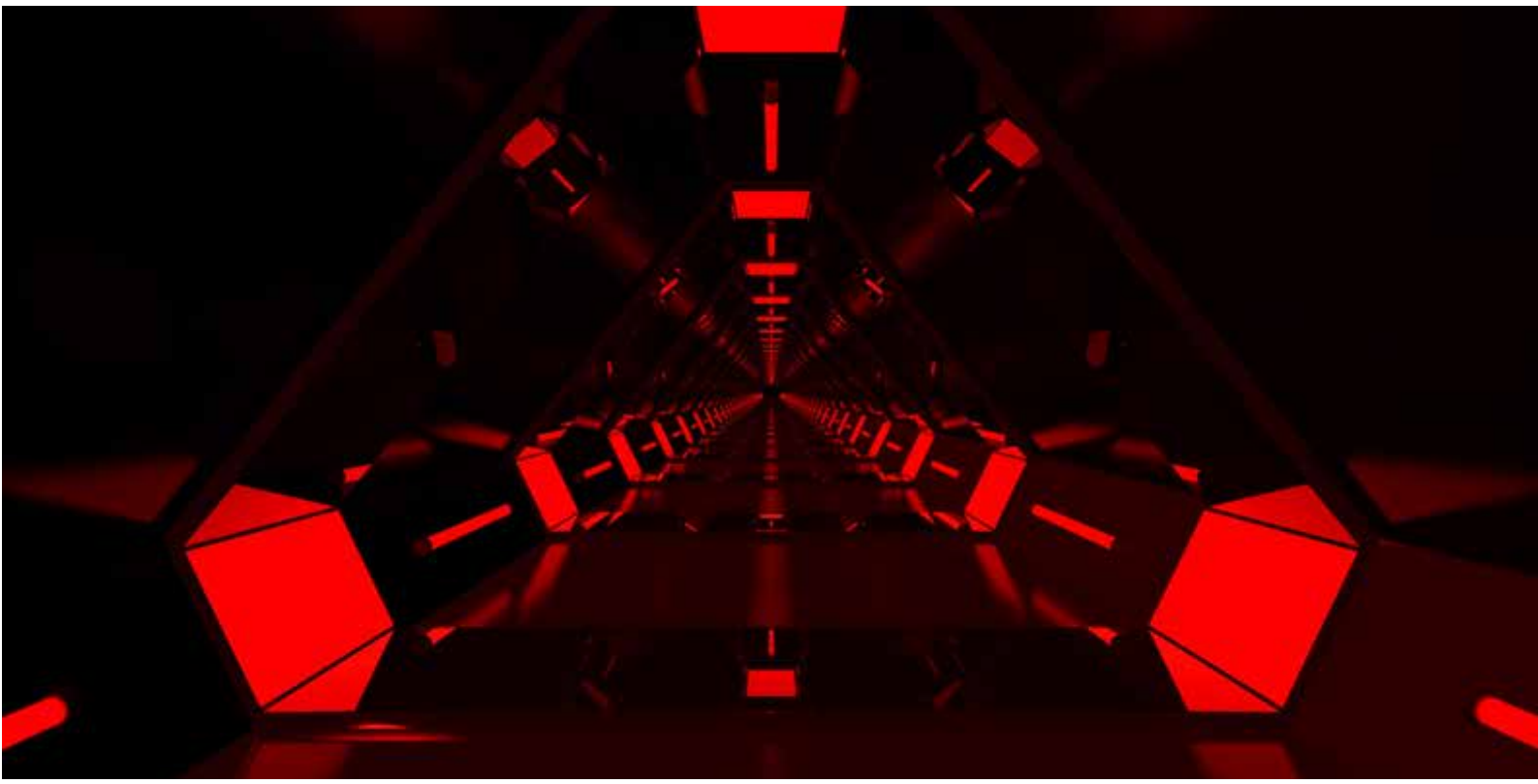
Lastly, validation is crucial. It's essential to evaluate if the solution benefits the business unit before scaling it up. This approach has proven effective for us."

Preeti SP, Digtial Technology Director at GE Appliances

Chapter 2

Cost Analysis

The cost analysis section will comprise several use cases in which the costs of using an external API will be compared with those of self-hosting on cloud GPUs. This section will also explore potential areas for cost optimization.



First Step to Tech Adoption: Calculating RoI

Implementing AI can be expensive, especially with large scale or high-accuracy requirements. It should ultimately bring a positive return on investment (ROI). This involves validating the AI's effectiveness with business units and being prepared to invest if it generates substantial business value.

In deploying Large Language Models (LLMs) within the MarTech sphere, a vital consideration is understanding and quantifying the ROI to evaluate the financial viability and success of this integration. Calculating the ROI involves comparing the net profit gained from utilizing LLMs to the initial and ongoing investment costs. The formula to calculate ROI is given by:

$$ROI = \left(\frac{\text{Net Profit from LLM}}{\text{Total Cost of LLM Implementation}} \right) \times 100\%$$

To break it down:

Net Profit from LLM: This represents the additional profits generated through the use of LLMs. It can be determined by evaluating metrics such as increased sales, enhanced customer engagement, and other revenue-generating outcomes attributed to the LLMs.

Total Cost of LLM Implementation: This encompasses the initial setup costs (like acquisition and integration), along with operational expenses, including maintenance and personnel training.

By applying this formula, organizations can visualize the profitability and efficiency of their investment in LLMs, thereby allowing for informed decision-making and strategy optimization in the MarTech sector.

The current section aims to give a guesstimate of the total cost of LLM Implementation and options to be explored based on the use case.

Using External APIs

Use Case 1 - Customer Complaint Redressal in the Automotive Industry

For a use case such as this, finetuning an AI model is crucial due to several reasons: the industry-specific terminology, unique automotive contexts, and the critical safety implications associated with potential misunderstandings. Finetuning ensures that the AI model offers precise, contextually relevant responses, minimizes manual oversight, and provides a competitive advantage by delivering a tailored customer experience.

Here is an estimate of the total cost for such a use case can be made:

Step - 1: Calculate the number of Instruction Set Tokens to Fine-tune the model

- Assume the number of customer complaints in the training dataset to around 6 million (6M complaints * 10 sentences/complaint * 7 tokens/sentence = 420 million tokens)
- Assume your corpus covers 20% of the text volume. Domain-Specific Corpus Size: 420M tokens * 0.2 = 84 million tokens
- Assume your instruction set tokens are proportional to the domain-specific corpus size and represent 50% of the corpus. Instruction Set Tokens: 84M tokens * 0.5 = **42 million tokens**

This cost is fixed and consists of a fine-tuning job with a training file of 42 million tokens that is trained for one epoch. The more specialized the size of the corpus, the better it will respond to queries. Of course, for a generic use case, it would mean fine-tuning on an entire corpus.

Step - 2: Calculate the number of tokens generated for Inference

Assuming that 3% of vehicle owners (150,000 owners) interact with the chatbot each month and each interaction involves an average of 1200 tokens (input and output), we come to the total number of tokens as approximately 150,000 * 1200 = 180 million tokens.

This cost is **recurring** and organizations have to bear every month based on the usage.

Step 3

Calculate the expected cost of Fine-tuning and Inference

Fine-tuning Models	Training	Training File	Fine-tuning Cost for 1 Epoch
babbage-002	$0.0004 / 1K tokens	42 million tokens	$17
davinci-002	$0.0060 / 1K tokens	42 million tokens	$252
GPT-3.5 Turbo	$0.0080 / 1K tokens	42 million tokens	$336

Fine-tuning Models	Input usage	Output usage	Input cost for 60 million tokens	Output cost for 120 million tokens	Total cost of inference for 1 month
babbage-002	$0.0016 / 1K tokens	$0.0016 / 1K tokens	$96	$192	$288
davinci-002	$0.0120 / 1K tokens	$0.0120 / 1K tokens	$720	$1,440	$2160
GPT-3.5 Turbo	$0.0120 / 1K tokens	$0.0160 / 1K tokens	$720	$1,920	$2640

babbage-002 and davinci-002 serve as replacements for the original GPT-3 base models and are trained with supervised fine-tuning. Meanwhile, GPT-3.5 Turbo is the most recommended choice due to its incorporation of RLHF.

Using External APIs

Use Case 2 - Sentiment Analysis for an eCommerce Company

Given the model's broad base training, it already understands diverse sentiments expressed online. Using **function calling in API**, users can extract organizational data within the prompt through an external database. This approach is not only cost-effective and versatile, but also allows for real-time customized responses to queries.

Here, the functions are billed as input tokens against the model's context limit.

Step - 1: Calculate the number of Input and Output Tokens

- Assuming there are about 1000 queries asked per day (30000 per month) to the model, averaging about 150 tokens per query for input and 500 tokens per query for output, we come to the total number as **5 million input tokens** and **16 million output tokens**.
- Assuming that each prompt carries a function call for an extra 150 input tokens, the total number would come to **10 million input tokens**.

This cost is **recurring** and organizations have to bear every month based on the usage.

Step - 2

Calculate the expected cost of Inference

Model	Input usage	Output usage	Input cost for 10 million tokens	Output cost for 16 million tokens
GPT-3.5 Turbo (4K context)	$0.0015 / 1K tokens	$0.002 / 1K tokens	$15	$32
GPT-3.5 Turbo (16K context)	$0.003 / 1K tokens	$0.004 / 1K tokens	$30	$64
GPT-4 (8K context)	$0.03 / 1K tokens	$0.06 / 1K tokens	$300	$960
GPT-4 (32K context)	$0.06 / 1K tokens	$0.12 / 1K tokens	$600	$1,920
Claude Instant (100K context)	$0.00163 / 1K tokens	$0.00551 / 1K tokens	$16	$88
Claude 2 (100K context)	$0.01102 / 1K tokens	$0.03268 / 1K tokens	$110	$523
Cohere (default)	$0.015 / 1K tokens	$0.015 / 1K tokens	$150	$240
Cohere (custom)	$0.03 / 1K tokens	$0.03 / 1K tokens	$480	$780

Low cost, better performance

medium cost, higher context length

Using External APIs

Optimizing Cost Using Embedding Models

Using pre-trained models from providers like OpenAI, textual data can be transformed into numerical vectors, known as embeddings. These models, typically based on neural networks, analyze the input data (like text) and represent it as a high-dimensional vector that captures its semantic features. After creating the embeddings, one can store them in a specialized vector database and perform queries to retrieve similar vectors or perform other analysis.

For a domain-specific corpus size of about **84 million tokens (50k documents)**, the following is the estimate cost:

When using embedding models, the vector database cost comprises most of the total cost.

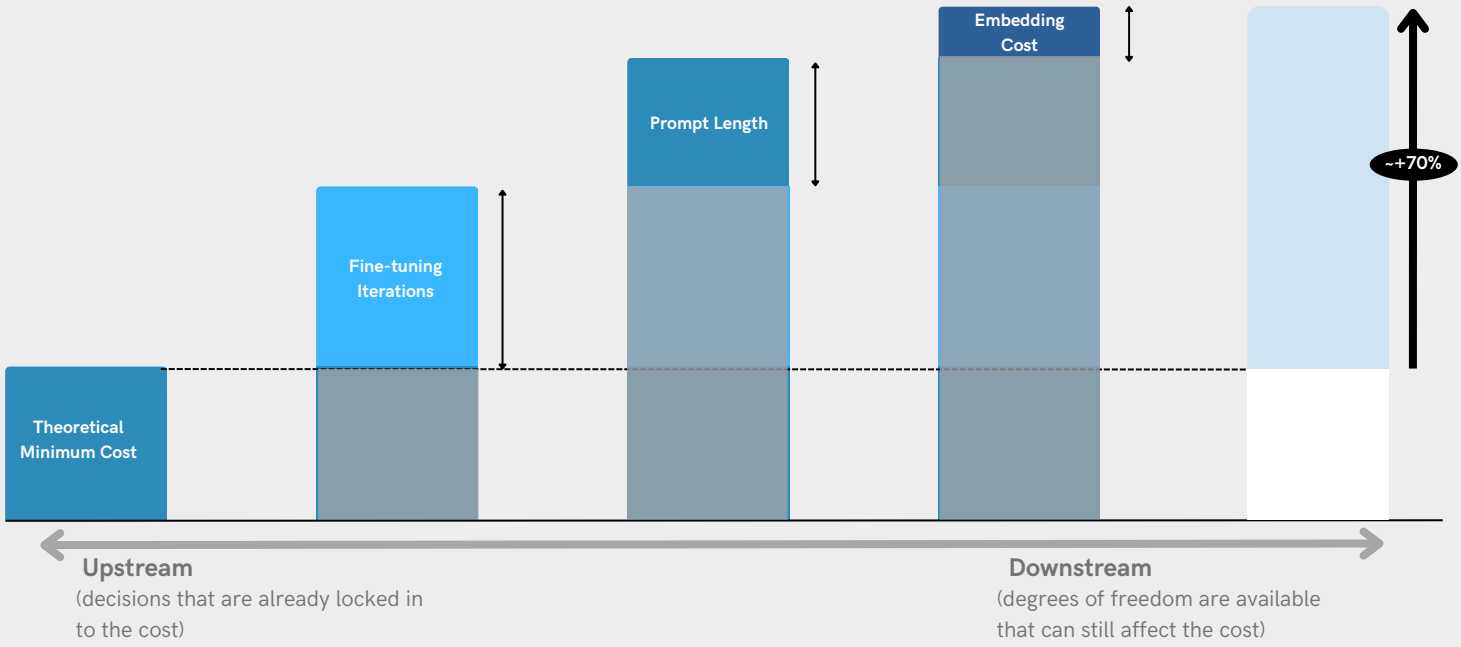
	Usage	Embedding dimensions	Usage cost per month	Vector DB monthly cost for 150QPS*
Ada v2	$0.0001 / 1K tokens	1536	$8	$400-$500
embed-english-light-v2.0	$0.0004 / 1K tokens	1024	$34	$250-$350
embed-english-v2.0	$0.0004 / 1K tokens	4096	$34	$1,500-$2,000

**This is just an average estimate; the cost will vary based on the kind of database used*

Using External APIs

Balancing Costs in API Integration

Companies are increasingly adopting generative AI APIs to get-to-market faster. However, crucial decisions arise: to fine-tune the AI or use it as is, and which prompt engineering techniques to apply. Beyond the initial investment, organizations should anticipate additional costs.



Self-hosting LLMs

Use Case 3 - Churn Analysis and Incentive-Based Win-Back in BFSI

In this use case, fine-tuning is essential because the sector possesses a unique lexicon and nuanced customer behaviors. By adapting models to BFSI-specific data, firms ensure higher accuracy in identifying churn patterns and crafting tailored incentives. Without fine-tuning, generic models might miss subtle sector-specific indicators, leading to less effective win-back strategies and potential revenue loss.

Here is an estimate of the total cost for such a use case can be made:

Step - 1: Calculate the number of Instruction Set Tokens to Fine-tune the model

- Assume the analyzed data comes from 5 million customers (5M customers * 100 sentences/customer * 7 tokens/sentence = 3.5 billion tokens)
- Assume your corpus covers 6% of the text volume. Domain-Specific Corpus Size: 3.5 billion tokens * 0.06 = 210 million tokens
- Assume your instruction set tokens are proportional to the domain-specific corpus size and represent 20% of the corpus. Instruction Set Tokens: 700 million tokens * 0.3 = **42 million tokens**

This cost is fixed and consists of a fine-tuning job with a training file of 42 million tokens that is trained for one epoch. The more specialized the size of the corpus, the better it will respond to queries. Of course, for a generic use case, it would mean fine-tuning on an entire corpus.

Step - 2: Calculate the number of tokens generated for Inference

Assuming that data is being analyzed for 5 million customers every month, and the analysis is performed twice a month. Furthermore, each analysis uses approximately 18 tokens per customer. This would bring the total to **180 million tokens** per month.

This cost is **recurring** and organizations have to bear every month based on the usage.

Step - 3

Calculate the expected cost of Fine-tuning and Inference

Model	Parameters	Compute (in TFLOPs)*	Configuration	Peak performance (in TFLOPs)**	Time (in hours)	Cost on AWS (for *one epoch*)
Llama 2	70 billion	1764 * 10^4	p4d.24xlarge	73	67.12	$2,199.63
Alpaca	65 billion	1638 * 10^4	p4d.24xlarge	73	62.33	$2,042.51
Falcon	40 billion	1008 * 10^4	p4d.24xlarge	73	38.36	$1,256.93
Vicuna	33 billion	831.6 * 10^4	p4d.24xlarge	73	31.64	$1,036.97

**Calculated as 6 x (number of tokens) x (number of parameters)*
***peak performance measured in a mixed precision setting*

Model	Parameters	Compute (in TFLOPs)*	Configuration	Peak performance (in TFLOPs)**	Time (in hours)	Cost on AWS (for *one epoch*)
Llama 2	70 billion	2520 * 10^4	p4d.24xlarge	73	95.89	$3,142.33
Alpaca	65 billion	2340 * 10^4	p4d.24xlarge	73	89.04	$2,917.88
Falcon	40 billion	1440 * 10^4	p4d.24xlarge	73	54.79	$1,795.62
Vicuna	33 billion	1188 * 10^4	p4d.24xlarge	73	45.21	$1,481.38

**Calculated as 2 x (number of tokens) x (number of parameters)*
***peak performance measured in a mixed precision setting*

The cost estimate is the maximum you'll incur under this configuration for the models considered.

Model	Parameters	Compute (in TFLOPs)*	Configuration	Peak performance (in TFLOPs)**	Time (in hours)	Cost on AWS (for *one epoch*)
Llama 2	70 billion	2520 * 10^4	p4d.24xlarge	73	95.89	$3,142.33
Alpaca	65 billion	2340 * 10^4	p4d.24xlarge	73	89.04	$2,917.88

In multi-GPU configurations, peak performance will increase. While this might lead to a reduction in cost, the performance boost almost never scale proportionally and is influenced by the level of optimization.

The number of tokens assumed for the sake of calculation is quite modest. In reality, many more tokens are generated for such a use case. This also implies that the costs will increase further. However, by using **reserved instances** for one or three years, enterprises receive a significant upfront discount, thereby reducing costs.

Fine-tuning reduces cost significantly

A crucial aspect of fine-tuning is the number of epochs required to refine the model effectively, directly influencing the cost. With self-hosting, the number of iterations decreases. Unlike an API, which often acts as a black box, open source offers insights into the model's inner hyperparameters.

But, there is a catch...

This advantage is somewhat offset by the cost associated with the many calls made during inference, which produce responses based on the provided sample answers. Moreover, maintaining a self-hosted model also incurs security-related expenses and other associated costs.

"Whether doing fine-tuning or one-shot/few-shot training, there's a tremendous amount of data and tokens involved. While these processes will eventually become economical, they are currently funded predominantly by large corporations. There are entities actively seeking ways to cut GPU costs and optimize algorithms. We need AI to be easily and cost-effectively customizable."

Nirupam Srivastava, Vice President - CX /AI, Legal and Startups Strategy at Hero Enterprise

Self-hosting LLMs

Use Case 4 - Content Personalization in FMCG

For FMCG, content personalization largely operates on generic, high-frequency data catering to broad consumer needs. Given the already extensive training of LLMs on diverse datasets, these models already possess knowledge relevant to FMCG sectors. Thus, instead of fine-tuning, effective prompting techniques can be employed when self-hosting, capitalizing on the model's inherent knowledge while ensuring agility and adaptability in generating personalized content.

Here, the functions are billed as input tokens against the model's context limit.

Step - 1: Calculate the number of Input and Output Tokens

- Assuming there are 5 personalized emails sent monthly, requiring 4 tokens per email for input and 22 for output, totaling 20 input tokens and 110 output tokens per user per month. For a relevant audience of 5 million users, email generation involves 100 million input tokens and **550 million output tokens**.
- During inference, the model generates two sampled responses, which brings the total to **1.1 billion output tokens**.

This cost is **recurring** and organizations have to bear every month based on the usage.

Step - 2

Calculate the expected cost of Inference

Model	Parameters	Compute (in TFLOPs)*	Configuration	Peak performance (in TFLOPs)**	Time (in hours)	Cost on AWS (for *one epoch*)
Llama 2	70 billion	1680 * 10^5	p3dn.24xlarge	125	373.33	$11,654.72
Alpaca	65 billion	1560 * 10^5	p3dn.24xlarge	125	346.67	$10,822.24
Falcon	40 billion	960 * 10^5	p3dn.24xlarge	125	213.33	$6,659.84
Vicuna	33 billion	792 * 10^5	p3dn.24xlarge	125	176.00	$5,494.37

Generally, the strategy is to start with a smaller configuration and gradually scale up to determine how you can achieve optimum results at the lowest cost. The cost reflected here is therefore on the higher side.

**Calculated as 2 x (number of tokens) x (number of parameters)*
***peak performance measured in a mixed precision setting*

Self-hosting LLMs

Optimizing Cost Using Optimization Libraries

LLMs require powerful computing clusters for training, but communication between GPUs can slow down the process and reduce efficiency. Proper management of this communication is crucial to avoid performance issues.

Distributed-training libraries available in the market introduce several management techniques that can improve the performance of GPU clusters. For example, Microsoft's Deepspeed uses a technique called the Zero Redundancy Optimizer (ZeRO) to achieve ideal scaling performance. In measurements, the RoBERTa-10B model running on AWS p4d.24xlarge instances achieved a performance of **123 teraflops per GPU** when optimized through Deepspeed, compared to only **73 teraflops** without optimization.

Model	Parameters	Compute (in TFLOPs)*	Configuration	Peak performance (in TFLOPs)**	Time (in hours)	Cost on AWS (for *one epoch*)
RoBERTa (without optimization)	10 billion	360 * 10^4	p4d.24xlarge	73	13.70	$448.90
RoBERTa (with optimization)	10 billion	360 * 10^4	p4d.24xlarge	123	8.13	$266.42

40%

**Calculated as 2 x (number of tokens) x (number of parameters)*
***as per data from Amazon*

Self-hosting LLMs

Balancing Costs in Self-hosting

Companies are progressively turning to self-hosted AI models for enhanced control and flexibility. However, critical choices arise: which hardware configuration to choose, whether to fine-tune or not, and which in-house optimization strategies to implement. Moreover, organizations must also be prepared for ongoing maintenance and operational expenses.

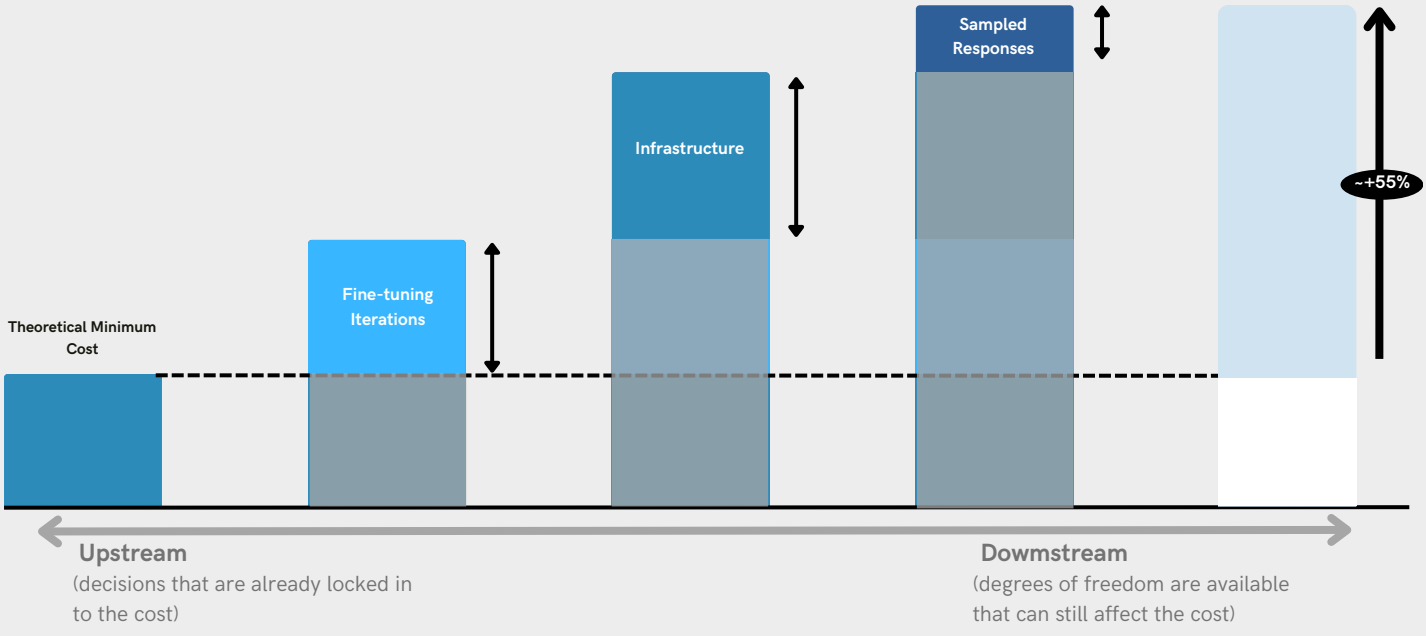
As emphasized earlier, the cost of fine-tuning decreases significantly with open-source models, as the number of iterations is reduced. Simultaneously, infrastructure costs can vary based on the performance/latency trade-offs that organizations choose.

"Beyond hardware costs, prompt creation is an overlooked challenge. It largely depends on an organization's dataset. While we are currently focused on data security, another issue emerges: siloed work. What about data fusion? Where is the comprehensive data lake designed with business needs in mind? Many fail to plan for or underestimate the associated costs."

Abhinandan Mandhana,
Executive Director, VP - Automation and Analytics at Bank of America

Exhibit 4

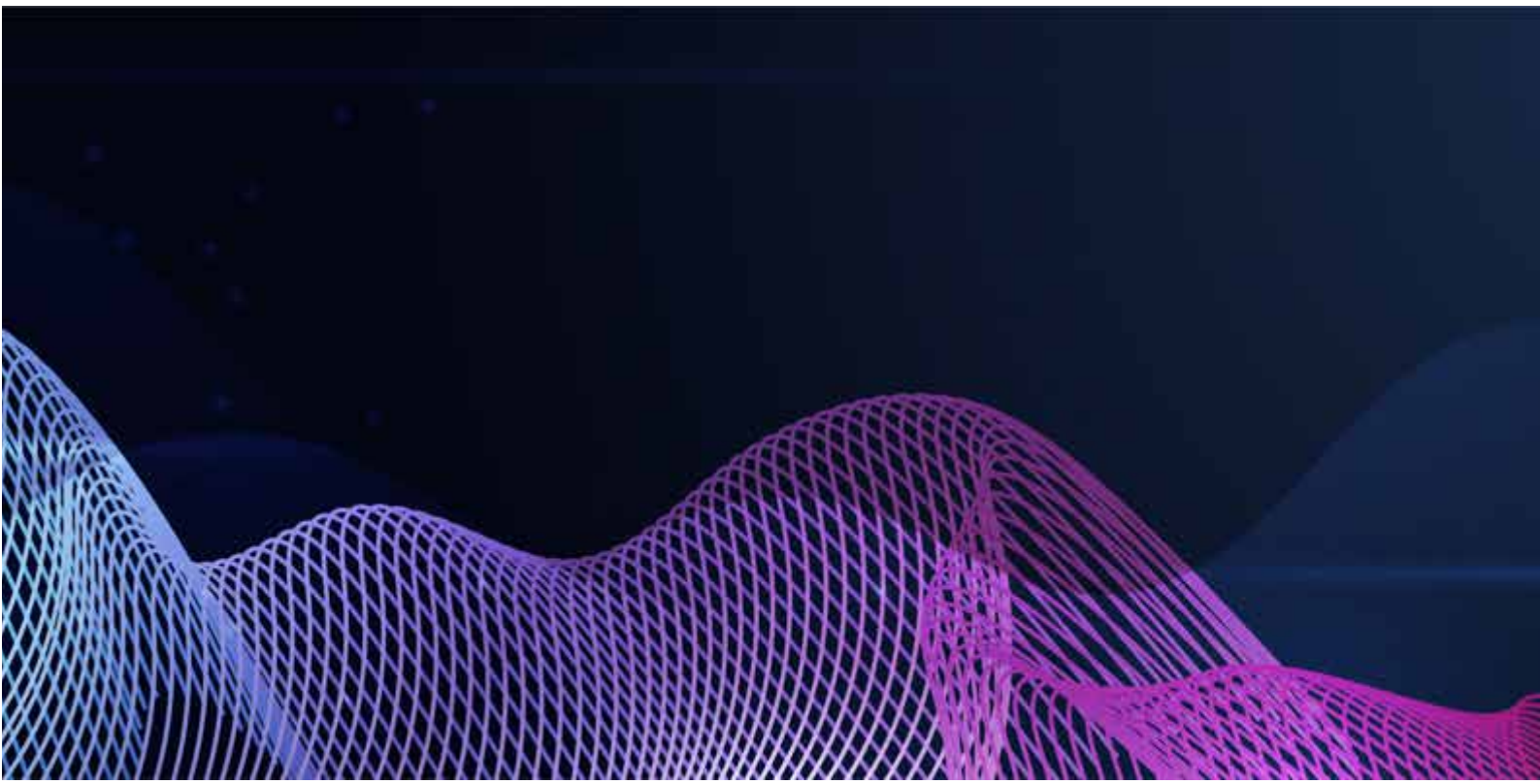
As the complexity of the use case increases, it will affect each of these cost variables to different degrees.



Chapter 3

How can you Reduce Cost?

This section will explore strategies for minimizing costs when utilizing external APIs and when self-hosting LLMs.



Industry is Moving Towards Cost-Effectiveness

In light of the recent observations on the high costs associated with AI, concerning both fine-tuning and one-shot and two-shot learning models, the current section on strategies to reduce costs in LLM implementation becomes exceedingly important. This section offering insights into cost-effective strategies and more sustainable approaches, hopes to pave the way for broader accessibility and adoption of AI technologies, without the burden of prohibitive costs.

Moreover, as we find the industry at a crossroads, with the present structures showing signs of strain under the monopoly of a few large corporations, the discussion on cost reduction gains even more relevance. This democratization of AI can potentially foster innovation and inclusivity in the sector.

Furthermore, considering the promising efficiency of narrower AI models, the guidance provided in this section could assist stakeholders in navigating the complex dynamics of AI implementation, aiding them in making informed decisions that balance both technological advancements and economic viability, thus facilitating a more sustainable and inclusive growth trajectory in the AI industry.

With External APIs

To reduce costs when using external API, organizations can implement various strategies:

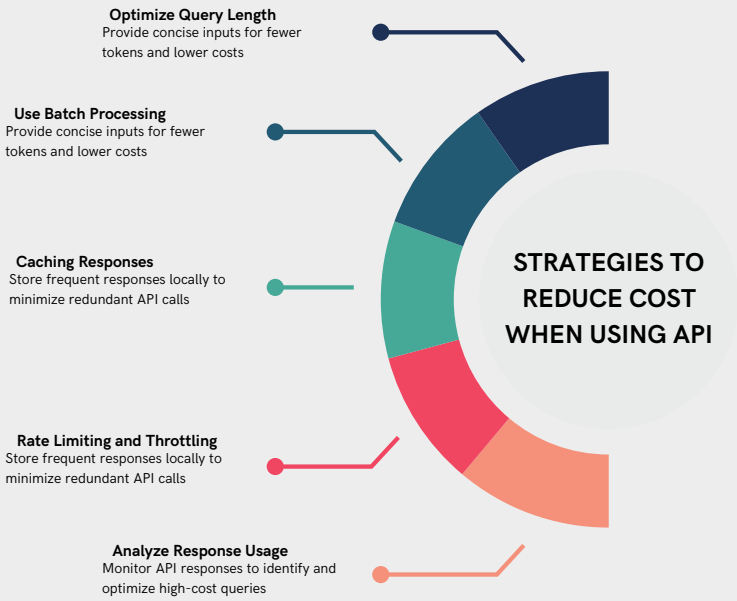
Optimize Query Length: Reduce the length of queries sent to the API by providing concise and relevant inputs. Shorter queries typically consume fewer tokens, resulting in reduced API usage and lower costs.

Use Batch Processing: Instead of sending individual queries one by one, batch multiple queries together and send them in a single API call. Batch processing can be more efficient and cost-effective as it optimizes API usage.

Caching Responses: Implement a caching mechanism to store frequently generated responses locally. By caching responses, the organization can minimize redundant API calls and lower usage costs.

Rate Limiting and Throttling: Apply rate limiting and throttling techniques to control the frequency of API requests. This prevents excessive API usage and helps manage costs effectively.

Analyze Response Usage: Monitor and analyze the responses generated by the API to identify patterns of high-cost queries. Optimize or modify these queries to reduce their impact on the overall cost.



With Self-hosted LLMs

To reduce costs when self-hosting LLMs, organizations can implement various strategies:

Optimize Hardware Resources: Choose hardware resources that match the model's requirements without overprovisioning. This involves selecting the right CPU, GPU, or TPU configurations based on the model's complexity and workload, minimizing unnecessary expenses.

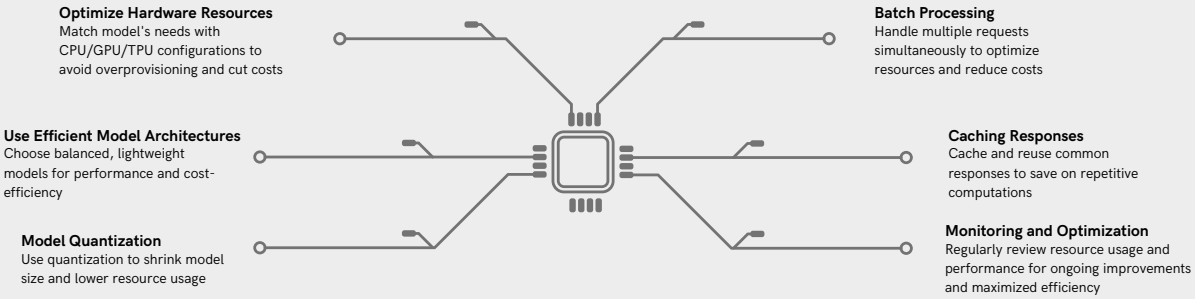
Use Efficient Model Architectures: Explore efficient model architectures that strike a balance between performance and resource consumption. Smaller or lightweight models can still deliver satisfactory results for many use cases while reducing computational costs.

Model Quantization: Apply quantization techniques to reduce model size and resource utilization. Quantized models use lower precision for computations, leading to memory and computational savings without significant loss in performance.

Batch Processing: Implement batch processing to handle multiple requests simultaneously, optimizing resource usage and reducing overhead costs.

Caching Responses: Employ caching mechanisms to store and reuse frequently generated responses, reducing the need for repetitive computations and lowering overall processing costs.

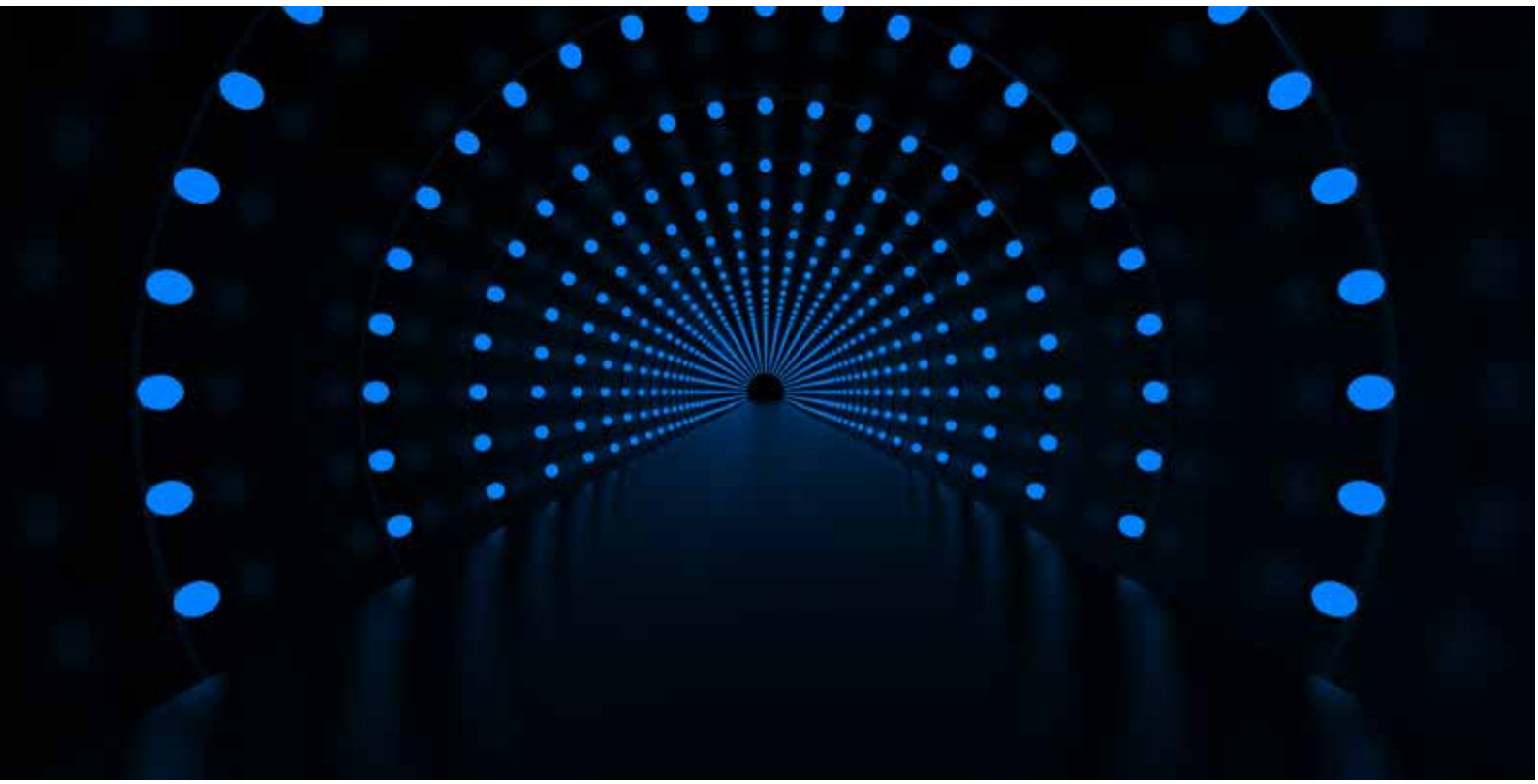
Monitoring and Optimization: Continuously monitor resource utilization and model performance to identify areas for improvement. Fine-tune the model and infrastructure configurations based on real-time data to maximize efficiency.



Chapter 4

Roadmap for Implementation

In this section, we outline a roadmap to financially sustainable implementations of text-based Large Language Models (LLMs). Focusing on long-term cost-effectiveness, we will explore strategies for optimizing expenses and resource allocation, ensuring a balanced budget while maintaining optimal performance.



Roadmap for a *Sustainable* Implementation of Text-based LLMs

The roadmap outlines the strategic framework for adopting and integrating language models into various workflows efficiently and sustainably. Emphasizing long-term value and scalability, the roadmap starts with a thorough assessment of current needs and future aspirations, ensuring the model aligns with the organization's goals.

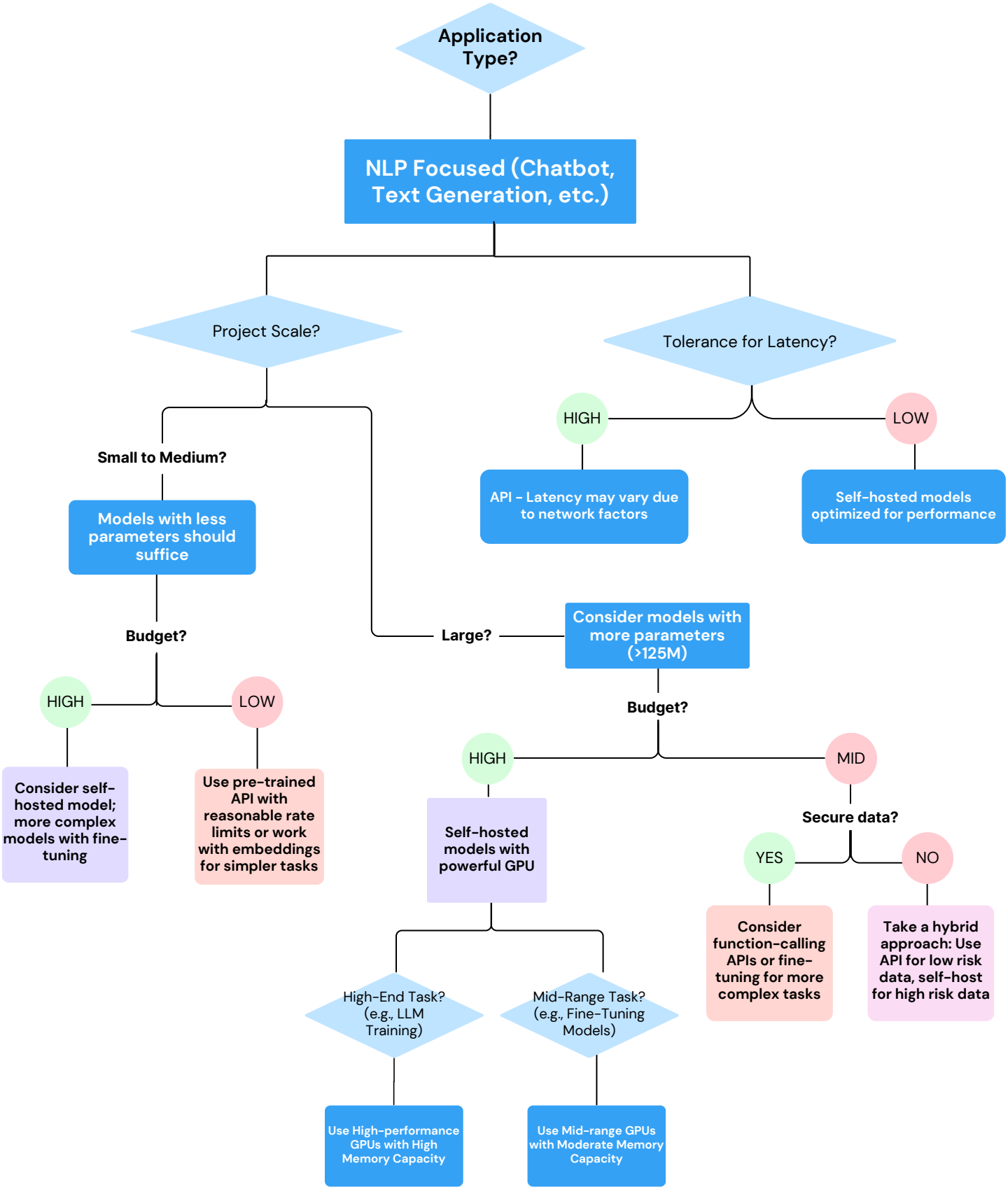
A crucial step not explicated here involves data management, ensuring high-quality, unbiased, and diverse datasets that can be used to train or fine-tune the model.

Moreover, given the rapid advancements in NLP, the roadmap will take into consideration continuous model updates and iterative learning. As implementation progresses, there's a focus on feedback loops, capturing real-world performance metrics, and adjusting accordingly.

"The real value emerges from how we integrate our organization's knowledge with large language models, and how we embed and connect these two aspects to provide applications for both internal and external customers. Readiness involves conducting an internal assessment: What is our data readiness? How is our data engineering infrastructure currently built? If we start now, how will we handle incremental data, and how often will we update it with experimental data? A thorough review of our data engineering approach is essential, considering how our data is collected, stored, and which data we intend to use."

Anand Mahalingam, Vice President - Data Labs- Head of AI at HDFC Life

Roadmap for a *Sustainable* Implementation of Text-based LLMs



Conclusion

In conclusion, as we step into a period that is expected to be quite dynamic in the coming years, it's crucial to keep a close eye on the changes happening in the marketing departments of various industries. These shifts, spurred by generative AI advancements, are setting the stage for deeper and more personalized connections with customers.

There is no one-stop solution to all use cases. It is better to experiment with a mix of both API and self-hosted models. For use cases that need to be shipped early to measure an early impact, implementation can be done via API, while those involving crucial organizational data can be implemented through different levels of experimentation in terms of data readiness, prompt engineering, fine-tuning, and hardware.

While the initial costs for self-hosting models are high, they can be more economical for large-scale use cases in the long run, with the ongoing operational costs being much lower than using an API. This approach also grants organizations more control over their data. Concurrently, new developments are emerging each day. For instance, ChatGPT Enterprise was recently launched by OpenAI, promising enhanced security for the internal data of organizations, thereby making this an attractive option for them.

What also affects the decision is the type of accuracy and latency you're targeting. While higher accuracy and workloads tend to increase costs, adjusting your hardware gives you an option to forgo a bit of latency for performance.

Looking ahead, it is clear that we are just at the beginning of this journey. The insights gathered here are only a starting point. Our upcoming reports will provide a more detailed analysis, backed by our research and findings, highlighting the ongoing developments in this new and exciting field. These new editions will serve as a reliable source of information, offering updates on the rapid changes and developments taking place in this promising field.

Acknowledgements

The preparation of this report was greatly enriched by the invaluable contributions of numerous professionals including **Narasimha Medeme**, VP Head Data Science at MakeMyTrip, **Ashwin Swarup**, VP Data Science at NimbleWork.Inc., **Anirban Nand**i, VP AI Products & Business Analytics at Rakuten India, **Sourav Banerjee**, Head of Innovation at TheMathCompany, and **Srinath Sivalenka**, Senior Manager - Generative AI Capabilities, who generously dedicated their time and expertise to the peer-review process.

To each individual named and the many others who contributed their time and expertise in various capacities, we extend our deepest appreciation. The candid discussions, recommendations for further study, and critical evaluations offered by our peers were paramount in refining our methodology and ensuring the accuracy of our conclusions.

About Hansa Cequity

Hansa Cequity is India's first data-driven marketing consulting & services company with a focus on Consulting, Data Management, Data Science, Behavioural Science, MarTech, Data-driven digital solutions and Customer Relationship Centres for different clients across key verticals like BFSI, Automotive, Media & Entertainment, Retail, Travel & Hospitality and E-Commerce. It is a part of the R K SWAMY Group, India's leading Integrated Marketing Communication services provider.

Hansa Cequity is a leader in India providing data-driven marketing solutions & services for blue-chip companies across India. It holds and analyses over 100 million unique customer profiles in private & public cloud infrastructure with more 100 terabytes of data & manage over 750 million one-to-one customer intelligence interactions in a year. Hansa Cequity has a team of more than 1000 consultants and associates in their key client engagements & programs.

ENRICHING CUSTOMER EQUITY

ISO/ IEC 27001:2013 CERTIFIED

Organisations across the world utilize us for advice and tools to lead their digital transformation using data.

Gain insights, advice and tools to embed analytics within your organisation. Equip yourself better to make decisions on AI capabilities

aimresearch.co

AIM India
#280, 2nd floor, 5th Main, 15 A cross, Sector 6, HSR layout Bengaluru, Karnataka 560102

AIM Americas
2955, 1603 Capitol Avenue, Suite 413A, Cheyenne, WY, Laramie, US, 82001

www.aimresearch.co

info@aimresearch.co