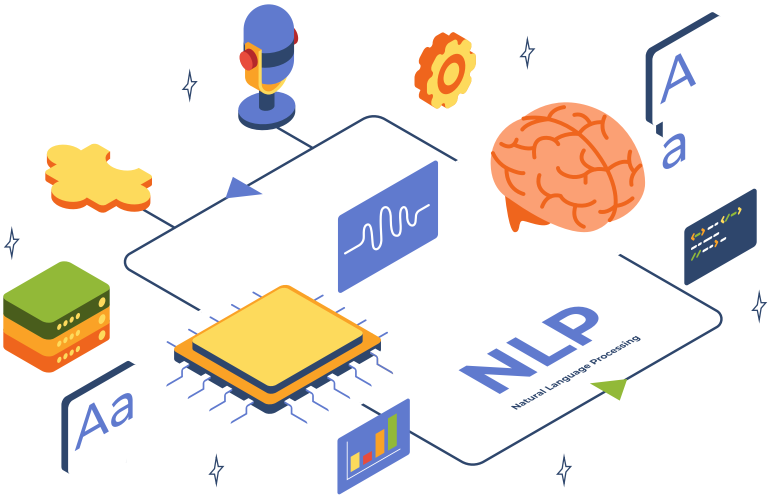
Making Generative AI Work

Choosing the Right LLM or SLM for Your Enterprise

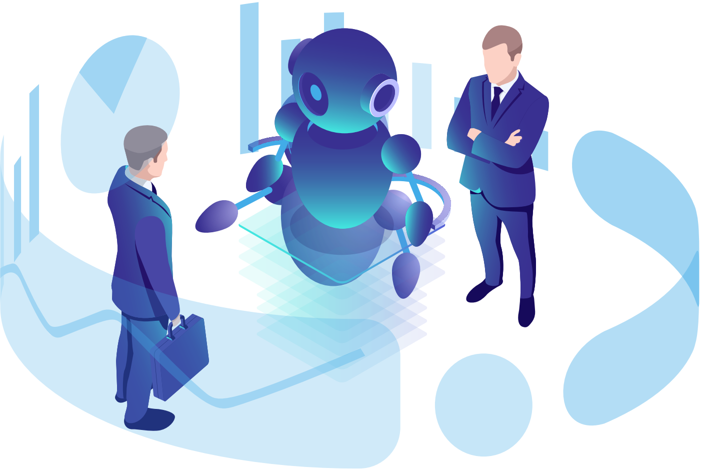
www.azilen.com

00 Introduction



Generative AI has moved from being a futuristic concept to a reality that businesses are expected to embrace. It's everywhere – from automating content creation to enhancing customer experiences. But as the hype around Generative AI grows, so does the pressure on companies to not just explore AI, but to actually make it work.

For product owners, this isn't just a matter of keeping up with the latest tech trends. They're feeling the heat from all sides. Customers want smarter products, markets demand innovation, and investors expect AI to drive growth. This collective pressure – what we refer to as "The Hype Pressure"– creates an urgent need for strategic planning in the adoption of Generative AI models.



01

The Real Challenge:

Making the Right Choices

> *Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept by End of 2025*

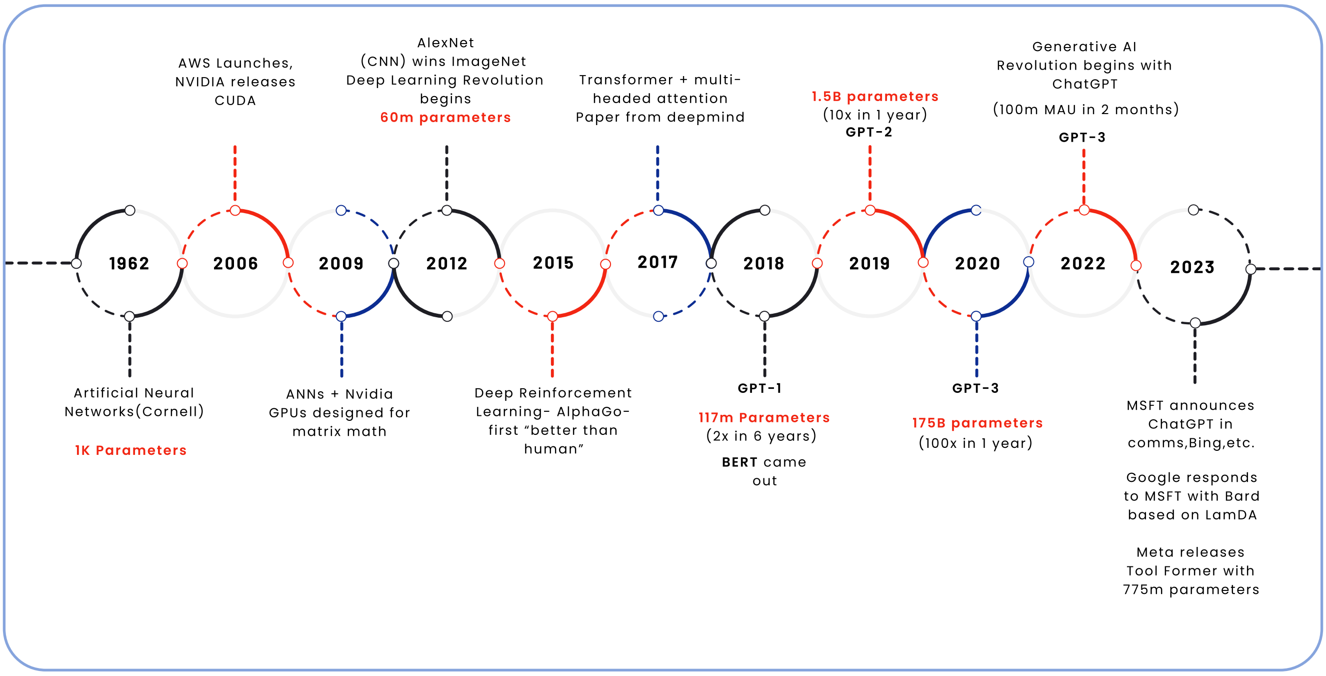
While the opportunities are vast, adopting Generative AI is not a straightforward process. The core challenge lies in making informed decisions about where and how to implement these models. Enterprises must grapple with critical questions: Which use cases are most viable for AI integration? How can we ensure that our approach aligns with our business objectives?

This white paper aims to unpack these complexities surrounding the adoption of Generative AI models in modern enterprises. We will guide product owners through the nuances of navigating the hype, addressing the challenges of selecting the right use cases and approaches, and showcasing real-world examples of organizations that have successfully embraced Generative AI — and those that have faced hurdles along the way.

02

The Rise and Shift in Generative AI and LLMs

Rise of Generative AI



1962

Artificial Neural Networks(Cornell)

1K Parameters

2006

AWS Launches, NVIDIA releases CUDA

2009

ANNs + Nvidia GPUs designed for matrix math

2012

AlexNet (CNN) wins ImageNet Deep Learning Revolution begins

60m parameters

2015

Deep Reinforcement Learning- AlphaGo- first "better than human"

2017

Transformer + multi-headed attention Paper from deepmind

2018

GPT-1

117m Parameters

(2x in 6 years)

BERT came out

2019

1.5B parameters

(10x in 1 year)

GPT-2

2020

GPT-3

175B parameters

(100x in 1 year)

2022

Generative AI Revolution begins with ChatGPT

(100m MAU in 2 months)

GPT-3

2023

MSFT announces ChatGPT in comms,Bing,etc.

Google responds to MSFT with Bard based on LamDA

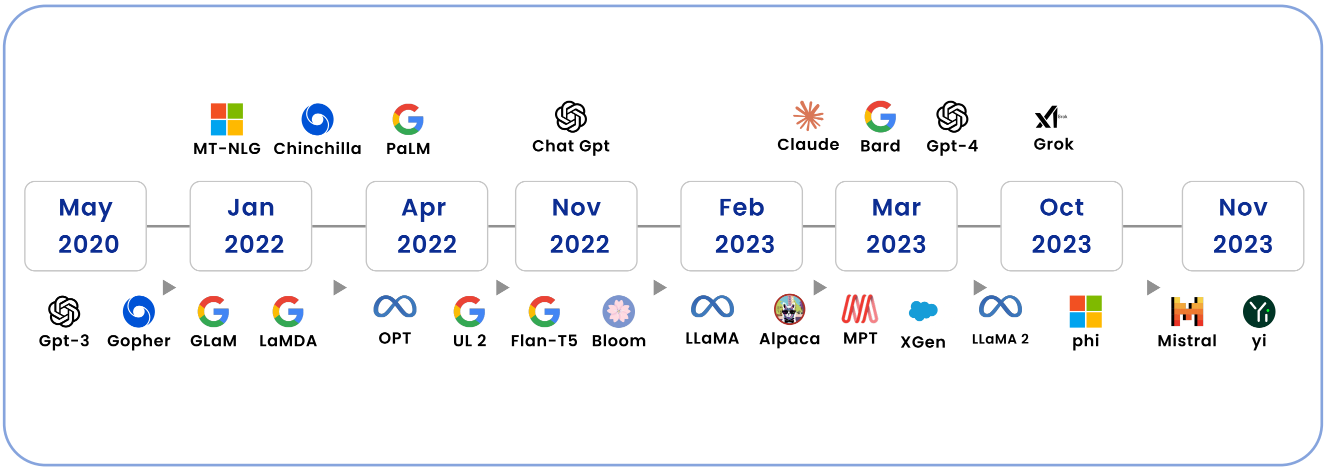
Meta releases Tool Former with 775m parameters

From its roots in rules-based systems, Generative AI now runs on powerful recurrent neural networks (RNNs), backed by robust hardware and software capabilities. Each incremental addition propels AI forward, placing us both at the beginning and in the midst of the Generative AI revolution.

ChatGPT marked a pivotal moment in November 2022. What was once considered cutting-edge in September 2022, is now already common, showcasing the rapid evolution of this technology. Corporations and Generative AI are already intertwined, and those connections will deepen even more as we embark on new breakthroughs shaping the future of innovation.

Rise of LLM

> *Almost 67% of organizations use generative AI products that rely on LLMs to work with human language and produce content.*

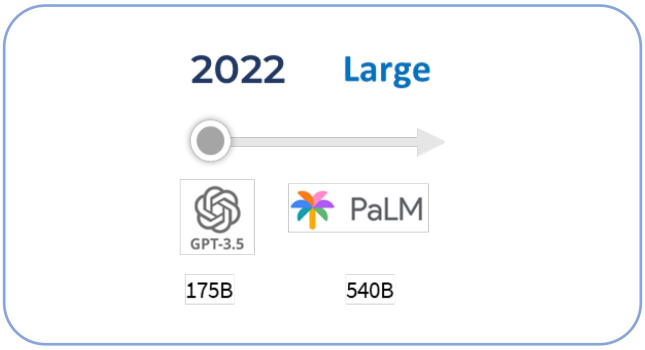


The release of GPT-3 in 2020 by OpenAI was a game changer for artificial intelligence and natural language processing. With an impressive 175 billion parameters, GPT-3 demonstrated an extraordinary ability to understand and generate text that feels incredibly human-like. This milestone opened up a world of possibilities, sparking a wave of interest and investment in Generative AI technologies across various industries.

> Fast forward to today, and organizations are increasingly embracing hybrid models that blend LLMs with other AI technologies, like computer vision and reinforcement learning. We're also seeing significant advancements in model architectures, which have made real-time interactivity a reality. With the emergence of models like GPT-4 and others that handle multimodal inputs — combining text, images, and audio — the user experience is richer than ever.

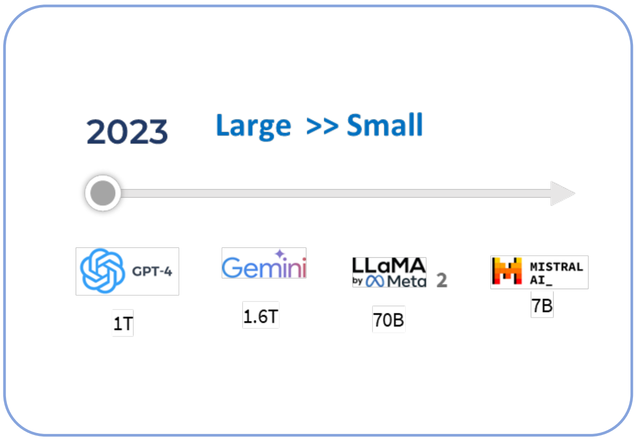
Shift in Model Sizes: From Large to Tiny

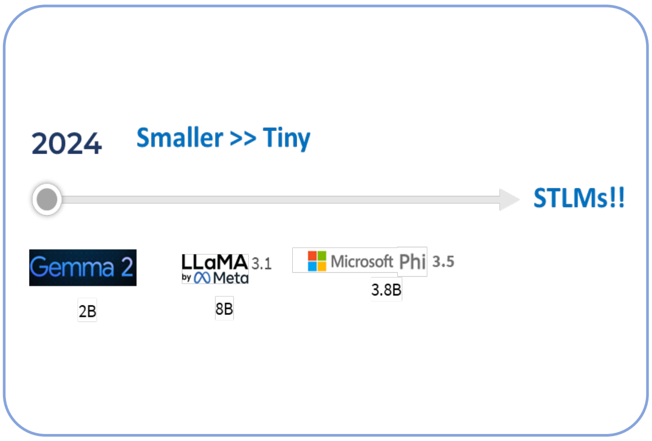
The shift from large models to smaller and even tiny models is exemplified by the following developments:



The landscape was dominated by large models, such as GPT-3.5 with 175 billion parameters and PaLM with 540 billion parameters. These models showcased significant capabilities in generating human-like text and performing complex tasks.

The trend began to pivot towards smaller models, exemplified by GPT-4 (estimated at 1 trillion parameters), Gemini (1.6 trillion parameters), LLaMA (70 billion parameters), and Mistral (7 billion parameters). This shift indicated a growing recognition of the advantages of smaller, more efficient models in terms of cost and performance.





The focus further narrowed to tiny models, with developments like Gemma (2 billion parameters), LLaMA 3.1 (8 billion parameters), and Microsoft Phi 3.5 (3.8 billion parameters) gaining traction. These smaller models are designed for specific tasks, providing effective solutions without the resource overhead associated with larger models.

03

Challenges with LLMs

As enterprises increasingly adopt LLMs, they face significant challenges related to training costs, energy consumption, and carbon emissions. These challenges underscore the complexity and environmental impact of training such large-scale models, which must be addressed to ensure their sustainable and efficient integration into enterprise ecosystems.

> *The global electricity consumption attributed to AI could surge by 85-134 TWh annually by 2027, equivalent to the annual energy usage of countries like the Netherlands or Sweden.*

1. Training Costs

Training Costs

GPT-3

One Training Session:
$2M

Training an LLM like GPT-3 is not only resource-intensive but also extremely expensive. A single training session for GPT-3 is estimated to cost $2 million. This high cost stems from the immense computational power required to process vast amounts of data and refine the model's accuracy.

For enterprises, this raises critical questions about the affordability of scaling AI solutions. While the promise of AI-driven productivity is substantial, companies must weigh the financial burden of training and maintaining these models.

2. Energy Consumption

Energy Consumption

GPT-3

One Training Session:
1287 MWh

The energy consumption associated with training large models like GPT-3 is another significant concern. One training session for GPT-3 consumes approximately 1,287 megawatt-hours (MWh) of electricity. To put this into perspective, this is equivalent to the amount of energy consumed by an average U.S. household over more than 120 years.

This immense energy demand highlights the strain that AI places on global energy resources, particularly as enterprises adopt AI at scale. If such models are continuously developed, trained, and fine-tuned, the cumulative energy consumption will rise dramatically, increasing the urgency of addressing AI's environmental and operational impacts.

3. Carbon Emissions

Carbon Emission

GPT-3

One Training Session:
502 Metric Tons CO2

In addition to high energy consumption, training large models like GPT-3 contributes significantly to carbon emissions. A single training session for GPT-3 results in the emission of 502 metric tons of CO$_2$, which is comparable to the annual emissions of 110 typical passenger vehicles or the energy used by an average household over 56 years.

This high carbon footprint underscores the environmental costs of AI development. With global efforts focused on reducing carbon emissions, the current trajectory of AI poses challenges to sustainability initiatives.

04

The Small is a New BIG

> *Google on TinyBERT said that it retains about 90% of BERT's performance while being 7.5 times smaller.*

While LLMs have dominated the conversation with their vast capabilities, Small Language Models (SLMs), with approximately 10 billion parameters or less, are quickly proving to be formidable in specific, targeted applications. This section explores what defines SLMs, their performance, and why they are increasingly important for modern enterprises.

What are SLMs?

SLMs are lightweight neural networks designed to handle natural language processing tasks with fewer parameters and lower computational demands than LLMs. While LLMs are often built to handle a wide range of tasks, SLMs are typically designed for specific purposes or use cases.

Parameters: Approximately 10 billion or less.
Computational Demands: Lower than those of LLMs.

Key Differences	LLMs	SLMs
Design Purpose	Generalized tasks	Specific purposes/use cases
Parameter Count	Typically over 10 billion	Approximately 10 billion or less
Efficiency	Broader applications	Highly specialized, efficient

The Small is new BIG - SLMs for the Usecases

Tasks

- Text Generation & Classification
- Sentiment Analysis
- Question Answering
- Language Translation
- Named Entity Recognition (NER)
- Intent Recognition

Models

- LLaMA 3.1 by Meta
- MISTRAL AI_
- VICUNA
- Gemma
- Microsoft Phi 3.5 mini

Value Propositions

- Computational Efficiency
- Deployment Flexibility
- Faster Training and Interface Times
- 40%-60% Smaller than LLM(s)
- Leverage Pruning for Model Optimization
- Lower Model Operational Costs

Performance of SLMs: Small but Mighty

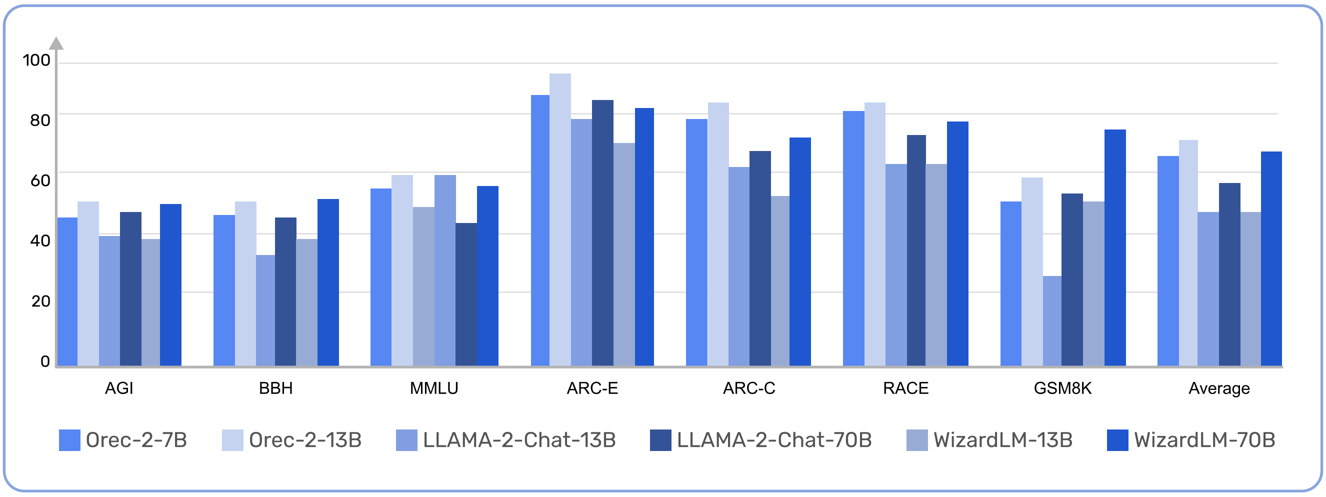
Recent research highlights the surprising performance capabilities of SLMs, often rivaling larger models in specific contexts. A prime example is phi-3-mini, an SLM with 3.8 billion parameters trained on 3.3 trillion tokens. Despite its size, the phi-3-mini's overall performance rivals that of much larger models, such as Mixtral 8x7B and GPT-3.5. In terms of standardized benchmarks, phi-3-mini achieved impressive results — 69% on the MMLU benchmark and 8.38 on the MT-bench — demonstrating its ability to handle complex tasks efficiently.

Additionally, phi-3-vision, a 4.2 billion parameter model derived from phi-3-mini, excels in handling multimodal inputs like image and text prompts, further expanding the utility of SLMs across various domains. Even earlier versions of SLMs, such as phi-2, have outperformed models up to 25 times larger in several complex tasks, reinforcing the notion that bigger isn't always better when it comes to AI models.

Another notable example is Orca2, designed specifically for research, which has showcased its ability to handle intricate tasks despite its smaller size.

Key Insights

- **phi-3-mini:** Rivals larger models (e.g., Mixtral 8x7B, GPT-3.5).
- **phi-3-vision:** Excels in multimodal inputs (image and text).
- **Previous Versions:** Even earlier SLMs like phi-2 have outperformed models up to 25 times larger in various complex tasks.



Results comparing Orca 2 (7B and 13B) to LLaMA-2-Chat (13B and 70B) and WizardLM (13B and 70B) on a variety of benchmarks (in zero-shot setting) covering language understanding, common-sense reasoning, multi-step reasoning, math problem solving, etc. Orca 2 models match or surpass other models, including models 5-10 times larger. Note that all models in this figure share the same base model (LLAMA-2).

The Rise of Ultra-Compact Models

The trend toward creating even smaller models continues, with models like TinyLlama and OpenELM leading the charge. TinyLlama, introduced in late 2023, features just 1 billion parameters but has set new benchmarks for ultra-compact models. Meanwhile, Apple's OpenELM, launched in April 2024, emphasizes the use of SLMs in edge devices, highlighting their ability to operate efficiently on local hardware without reliance on cloud-based processing.

The move towards these smaller, localized models is driven by the need for AI solutions that are not only performant but also more resource-efficient and environmentally sustainable.

Why SLMs Matter

1. Efficiency and Speed

One of the most significant advantages of SLMs is their efficiency. With fewer parameters to manage, SLMs are much faster to train and offer quicker inference speeds. This means that organizations can deploy AI-driven solutions more rapidly, allowing for faster time-to-market for applications.

2. Sustainability

SLMs contribute to a more sustainable AI ecosystem. Their reduced computational requirements translate to lower energy consumption during training and inference. This leads to smaller carbon footprints and less water usage, making SLMs an eco-friendly option compared to the resource-intensive LLMs. The environmental benefits align with growing corporate and societal demands for greener technology solutions.

3. Cost-Effectiveness

SLMs are significantly more cost-effective to train and deploy. With fewer parameters, the computational costs associated with SLMs are lower, making these models more accessible to companies that may not have the extensive resources required to develop or operate LLMs. Moreover, offloading processing tasks to edge devices reduces the need for expensive cloud infrastructure, lowering both capital and operating costs.

4. Targeted Performance

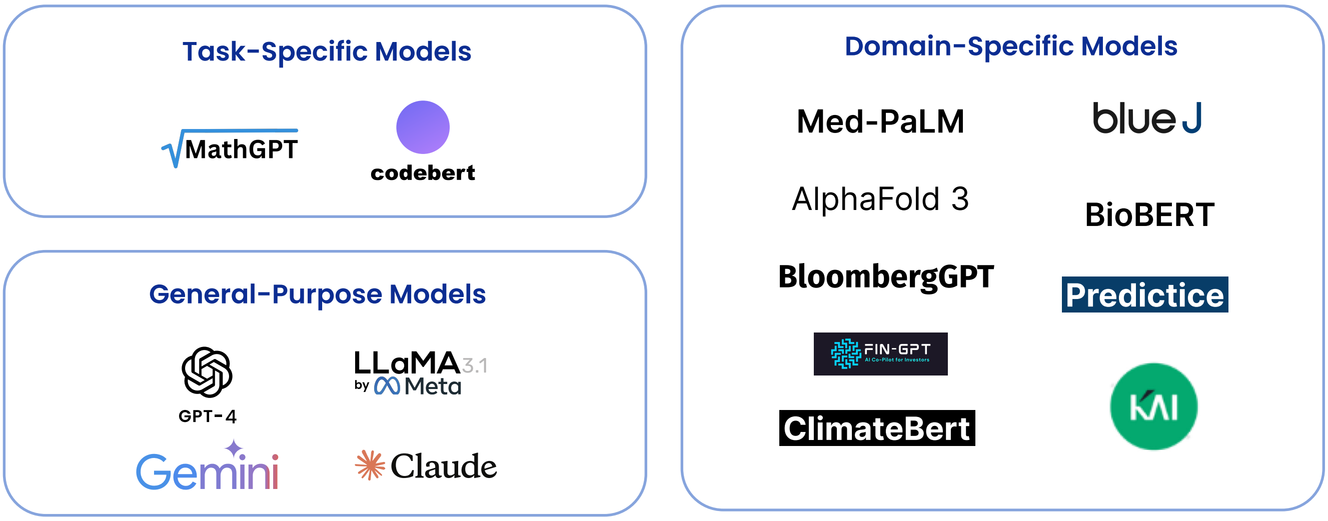
SLMs excel at delivering high performance in specific tasks. By focusing on particular use cases, SLMs can be fine-tuned to outperform larger, more generalized models in those domains. This tailored approach makes SLMs ideal for enterprises looking to implement AI solutions that address specific challenges without the overhead of maintaining massive models like GPT-4 or LLaMA.

05

From General-Purpose LLMs to Task and Purpose-Specific SLMs

While LLMs like ChatGPT, GPT-4, or Google's Gemini offer broad capabilities and can perform a wide array of tasks, the question remains: Do enterprises need such general-purpose models for every specific business problem? The answer increasingly points toward a no. As AI continues to mature, the move from general-purpose LLMs to task- and purpose-specific Specialized Language Models (SLMs) is becoming not just a trend, but a necessity in many industries.

From General-Purpose LLMs to Task & Purpose-Specific SLMs



Task-Specific Models

MathGPT

codebert

General-Purpose Models

GPT-4

LLaMA 3.1 by Meta

Gemini

Claude

Domain-Specific Models

Med-PaLM

blue J

AlphaFold 3

BioBERT

BloombergGPT

Predictice

FIN-GPT AI Co-Pilot for Investors

ClimateBert

KAI

The Case for Specialized Language Models

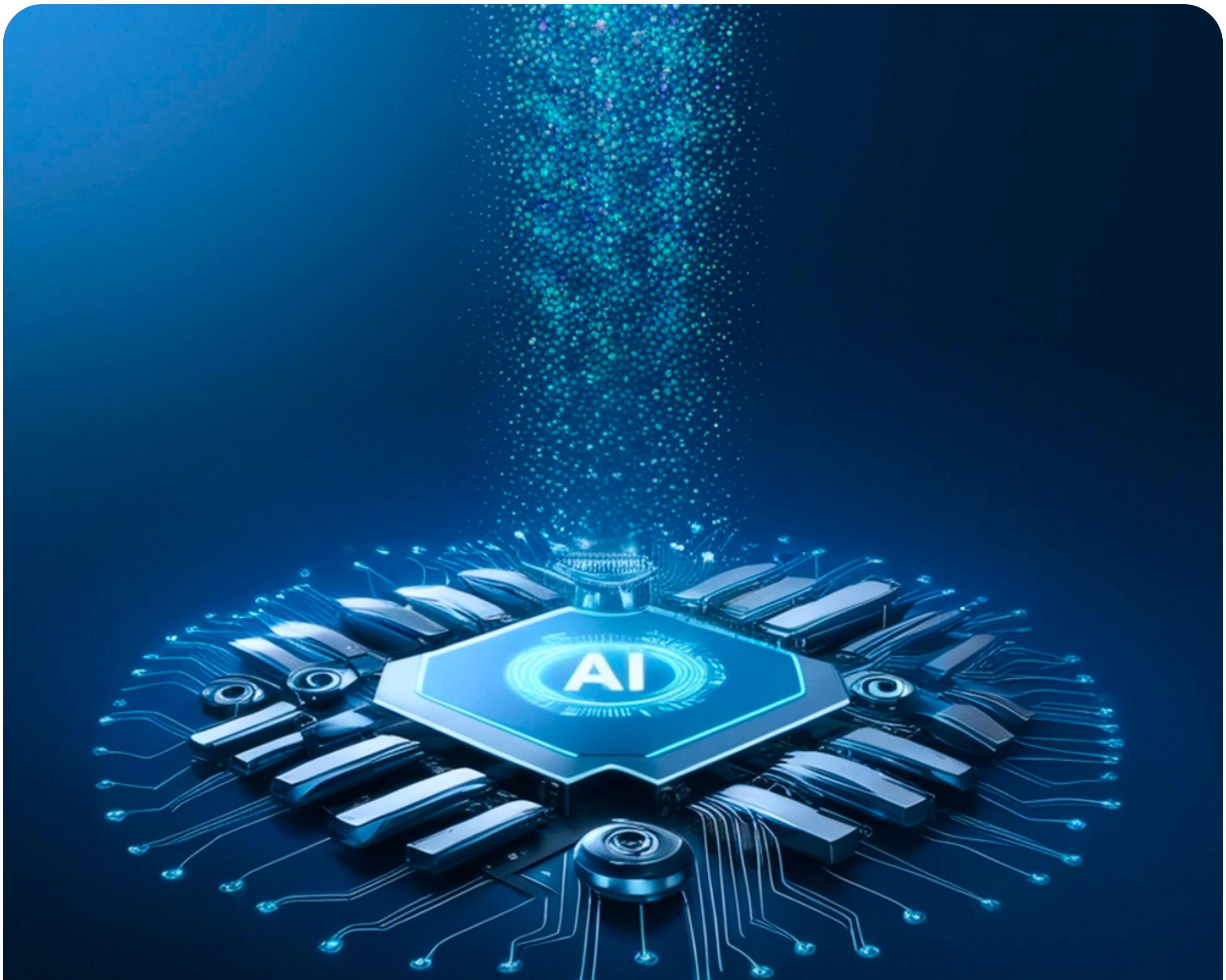
SLMs are tailored to excel in particular domains or tasks, offering precise, efficient solutions for specific business needs. These models are trained on data sets that are narrower in scope but richer in relevance to a particular domain, making them adept at solving specialized problems. Their specialization allows them to outperform general-purpose LLMs in specific applications where deep domain expertise is essential.

For instance, consider MathGPT, a model designed to solve advanced mathematical problems. Its training is focused on complex mathematical equations, algebraic functions, and other numerical tasks, enabling it to deliver accurate results that a general-purpose LLM may struggle to achieve with precision. Similarly, CodeBERT, trained specifically for understanding and generating code, outperforms models like ChatGPT when it comes to code interpretation, debugging, or software development tasks.

Key Features

- Targeted Training
- Efficiency
- Domain Expertise

Feature	Specialized Language Models (SLMs)	General-Purpose Language Models (LLMs)
Training Focus	Narrow domain-specific datasets	Broad and varied datasets
Performance	High accuracy in specific tasks	Good general performance
Use Cases	Advanced mathematical problems, coding	General conversations, writing, Q&A
Example Models	MathGPT, CodeBERT	ChatGPT, GPT-3
Domain Adaptability	Limited to specific domains	Versatile across multiple domains



Benefits of Task-Specific SLMs for Enterprises

Efficiency and Precision

SLMs are fine-tuned for targeted tasks, which makes them highly efficient in delivering relevant results. Whether it's understanding legal jargon, optimizing code, or analyzing scientific data, these models are more likely to produce accurate outputs in less time.

Resource Optimization

With a more focused scope, SLMs are often lighter and require fewer computational resources, making them easier to deploy and maintain compared to more complex LLMs.

Customizability

Enterprises can build or fine-tune their own SLMs based on proprietary datasets, creating models that align perfectly with their business needs. This customizability ensures that the AI is tailored specifically to solve their unique challenges.

Cost-Effectiveness

General-purpose models often require significant computing power and can be overkill for specific tasks. SLMs, by contrast, can be more cost-effective as they are optimized for specific use cases, reducing the need for extensive processing or excessive model tuning.

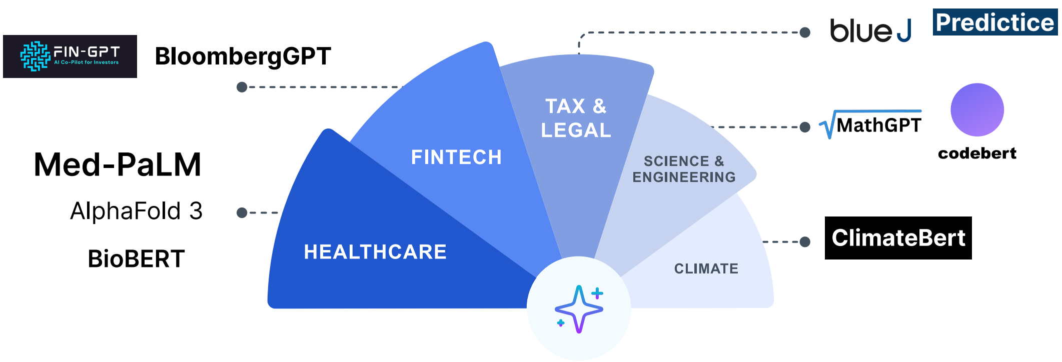
06

Domain-Specific LLMs for Different Industries

> *BioBERT outperformed BERT on biomedical tasks by 10-20%.*

While LLMs excel at understanding and generating broad-spectrum content, they may lack the precision, vocabulary, and nuanced knowledge required for more specialized domains like healthcare, finance, or law. This lack of specialized understanding can lead to inaccurate, vague, or irrelevant outputs in complex scenarios.

To address this gap, domain-specific LLMs have emerged as a refined solution. These models are trained on industry-specific datasets, focusing on specialized terminologies, processes, and contexts that are crucial for effective communication and decision-making within a given sector. The result is greater accuracy, relevance, and efficiency in both understanding and generating text, providing more reliable outputs that align with industry needs.



Enhancing Natural Language Understanding and Generation

Domain-specific LLMs enhance natural language understanding (NLU) by comprehending and responding to complex industry-specific queries, thus improving user interaction and engagement. In specialized sectors, this level of precision is crucial for building AI applications that deliver real value to users by providing answers that are not only technically accurate but also contextually meaningful.

Customer Service and Support

In the realm of customer service, domain-specific LLMs are game-changers. Customer service teams in industries like finance, healthcare, and retail deal with complex, specialized queries on a daily basis. General-purpose LLMs often struggle to provide precise responses to these queries due to the nuances of each industry. Domain-specific models, however, are trained on relevant customer service data, enabling them to offer personalized and accurate responses.

Healthcare and Biotechnology

The healthcare and biotechnology sectors, which rely heavily on precision and data-driven decisions, benefit immensely from domain-specific LLMs. These models can analyze large volumes of medical literature, patient data, and clinical studies, assisting healthcare professionals in making informed decisions regarding diagnosis, treatment options, and patient care.

In biotechnology, these models help researchers identify trends and correlations in data that can lead to innovative breakthroughs in areas such as genetics and molecular biology. The speed and accuracy with which these models operate enable researchers to stay ahead in highly competitive fields like personalized medicine and genomics.

Finance and Investment

In the highly competitive and data-driven world of finance and investment, domain-specific LLMs offer substantial value by processing and analyzing vast quantities of financial data. These models are adept at recognizing market patterns, identifying risks, and making predictions based on historical and real-time data.

Key Examples of Domain-Specific LLMs

BloombergGPT (Finance)

Developed by Bloomberg, this LLM is trained on vast amounts of financial data to provide market analysis, investment recommendations, and summaries of financial documents, improving efficiency in the financial sector.

FIN-GPT (Finance and Investment)

An open-source LLM focused on finance and investment. It uses Reinforcement Learning with real-time stock price data to provide insights and forecasts, assisting investors in making informed decisions.

Med-PaLM 2 (Healthcare)

Developed by Google, this healthcare-specific LLM is trained on medical datasets and excels at answering medical licensing exam questions, showcasing its potential to assist with clinical decision-making.

ClinicalBERT (Healthcare)

Built on the BERT architecture, ClinicalBERT is pre-trained on clinical texts to summarize patient data and predict readmission risks, improving patient management for healthcare professionals.

Blue J (Tax)

This Generative AI model designed specifically for tax experts, Blue J provides analysis and predictions on tax scenarios by reviewing past decisions, making legal research more efficient.

MathGPT (Mathematics)

This LLM is highly specialized in solving mathematical problems, outperforming general models like ChatGPT in accuracy and excelling in handling complex mathematical concepts.

GatorTronGPT (Healthcare)

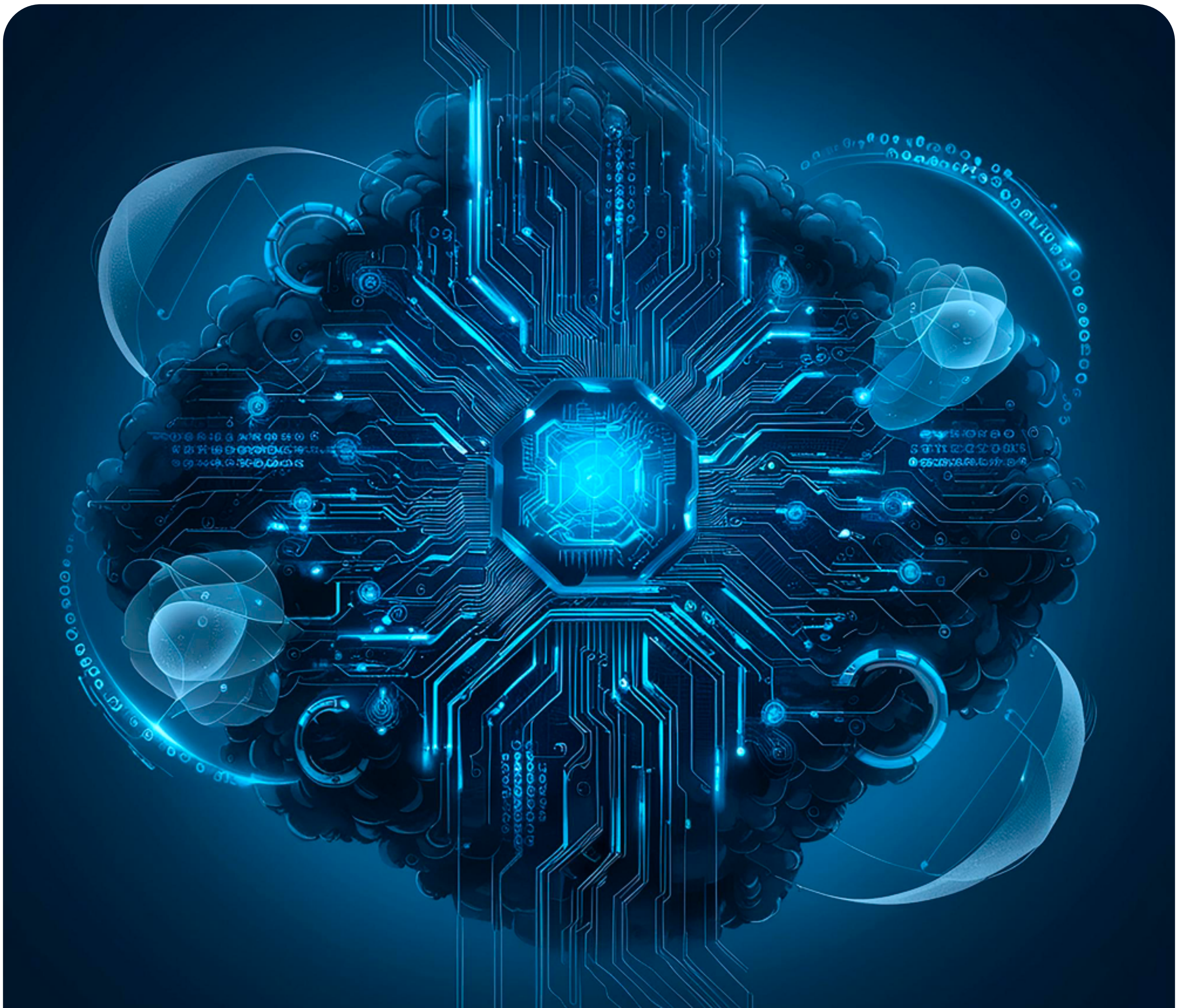
Developed by the University of Florida and NVIDIA, this model helps healthcare professionals by generating doctors' notes and assisting in processing large volumes of medical texts, enhancing decision-making in clinical settings.

BioGPT (Biomedical)

Pre-trained on 15 million PubMed abstracts, BioGPT supports biomedical text generation and mining, aiding researchers in drug discovery and scientific research.

KAI (Finance)

A conversational AI designed for the banking industry, powered by KAI-GPT. It provides precise answers for banking use cases, enhancing customer service and ensuring data security for financial institutions.



07

SLM vs. LLM – Cost Implications

The adoption of Generative AI in enterprises often hinges on decisions about which type of AI model to use — SLMs or LLMs. These choices are deeply influenced by cost considerations, as different models come with varying price tags that impact infrastructure, development, and operational expenditures. Let's explore the cost implications associated with both types of models, focusing on infrastructure, API pricing, and scalability.

1. API Usage and Cost

SLMs typically offer lower operational costs due to their smaller size and simpler architecture. These models can handle specialized tasks efficiently without the high computational demands of LLMs, making them ideal for enterprises with budget constraints or niche use cases.

For instance, OpenAI's API pricing shows that smaller models like GPT-3.5 are cheaper to use, charging a lower per-token rate. A lightweight model is often priced at a fraction of what larger models like GPT-4 might cost.

LLMs, by contrast, carry higher operational costs due to their size and complexity. OpenAI's GPT-4, Anthropic's Claude, and Google's Bard models are priced higher per token processed, reflecting the intensive computational resources required. A more sophisticated model like GPT-4 can cost more than 5x what a smaller language model would. (Know what is the cost of training LLMs)

SLMs	LLMs
Lower due to its smaller size and simpler architecture.	Higher due to size and complexity.
Lower per-token rate (e.g., GPT-3.5 cheaper).	Higher per-token rates (e.g., GPT-4 costs significantly more).
e.g., $0.03 per 1,000 tokens (GPT-4-8K)	e.g., $0.06 per 1,000 tokens (GPT-4-32K)

2. Infrastructure and Computational Costs

SLMs typically require less computational power and storage. This reduction in computational load also means enterprises save on cloud computing expenses like server time and energy consumption. Providers like Groq offer efficient pricing for hardware that supports smaller models with low latency and high-performance workloads, often resulting in a much more economical choice for enterprise applications that don't require massive processing power.

LLMs, on the other hand, are resource-intensive. Their training and deployment require vast amounts of GPU and TPU resources. Cloud providers like Azure charge premium rates for high-end GPU usage. For instance, LLMs like GPT-4 or Mistral's advanced AI models may require specialized hardware with high-throughput networking, leading to significantly higher cloud computing costs. Google Cloud and Azure AI charge by the second for high-performance VMs, and the cost can run into thousands of dollars per hour when deploying large-scale language models.

SLMs	LLMs
Requires less power and storage, reducing expenses	Resource-intensive, requiring vast GPU/TPU resources.
More economical; providers like Groq offer efficient pricing.	Premium rates for high-end GPU usage (e.g., Azure, Google Cloud).
Lower cloud computing expenses.	Costs can run into thousands of dollars per hour.

3. Scalability and Cost Trade-offs

Scalability is a major factor for enterprises evaluating AI models. As businesses scale up their AI usage, cost considerations become even more critical.

SLMs are often more affordable to scale, especially when their focus is on performing narrow, domain-specific tasks that don't require the vast knowledge base of an LLM. Enterprises might choose to deploy multiple SLM instances to cover different applications (e.g., customer service, fraud detection) without incurring high operational costs.

LLMs can provide multi-functional capabilities but at a much higher price. Their ability to handle broader and more complex tasks — like generating sophisticated reports, answering diverse queries, or summarizing large documents — makes them valuable but costly for enterprises aiming to scale.

SLMs	LLMs
More affordable to scale for narrow tasks (e.g., customer service)	Higher cost to scale for broader tasks.
Multiple instances for specific applications without high costs	Multi-functional but costly for enterprises aiming to scale.

4. Fine-tuning and Customization Costs

SLMs, being smaller and less complex, are generally easier and cheaper to fine-tune. Customization efforts, such as training the model on domain-specific data, require fewer computational resources. This can result in lower costs when compared to fine-tuning a larger model.

LLMs, on the other hand, present a more expensive fine-tuning process. The sheer size of these models means that fine-tuning requires significant computational power and time, leading to higher cloud costs. Additionally, the larger the model, the more data and compute cycles are needed to fine-tune it effectively, as seen in advanced models like Mistral AI's technology, which offers cutting-edge capabilities but demands more extensive resources.

SLMs	LLMs
Easier an chaper due to less complexity.	Expensive due to computational demands.
Requires fewer resources, leading to lower costs.	Larger models need more data and compute cycles, increasing costs.

5. Total Cost of Ownership (TCO)

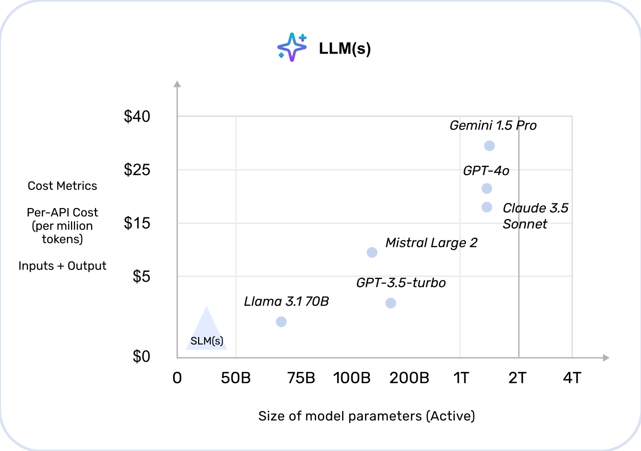
For many enterprises, the Total Cost of Ownership (TCO) goes beyond just the API or infrastructure pricing. It includes considerations like long-term maintenance, upgrades, and operational efficiency.

SLMs often have a lower TCO due to reduced maintenance complexity and fewer dependencies on high-end hardware or extensive data centers. This makes them attractive for smaller enterprises or those focusing on a few key tasks where a lightweight model can handle the workload.

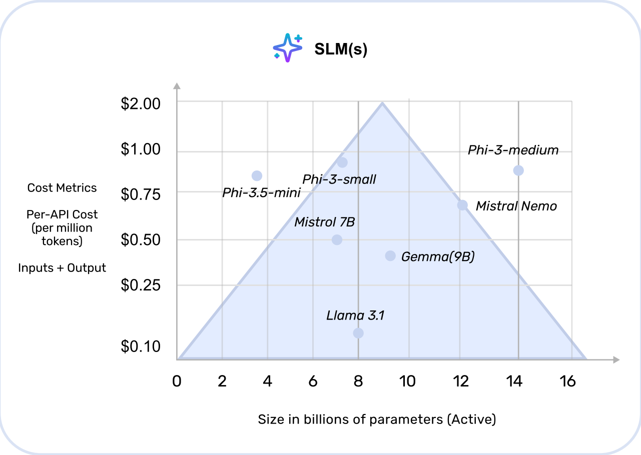
LLMs, due to their resource demands and operational complexity, tend to have a higher TCO. Enterprises must factor in ongoing costs related to infrastructure upgrades, potential downtime during fine-tuning or model retraining, and the need for robust data pipelines.

A Quick Look into SLMs vs. LLMs Cost Implications

LLMs - Cost Implications



SLMs - Cost Implications



08

LLMs and SLMs: The Right Approach for Adoption

As companies explore the potential of Generative AI, it's essential to grasp the differences between LLMs and SLMs. Both types leverage advanced machine learning techniques, but they serve different needs based on their design, training, and application.

Model Size and Complexity

The most apparent difference lies in size. LLMs, like ChatGPT (GPT-4), pack an impressive 1.76 trillion parameters. In contrast, smaller models such as Mistral 7B have around 7 billion parameters.

This size difference influences how these models learn and operate. LLMs use a sophisticated self-attention mechanism to understand context, making them versatile across various domains. Smaller models like Mistral 7B use a more streamlined approach, which allows for faster training and deployment but might not capture the same breadth of understanding.

Contextual Understanding and Domain Focus

SLMs excel in specific areas because they're trained on targeted data. This focus allows them to perform exceptionally well within their domain. However, they might miss out on the broader context.

On the other hand, LLMs aim for a wider understanding of language and can adapt to many tasks, making them great for diverse applications, from content generation to coding assistance. Their extensive training on varied datasets gives them an edge in versatility, even if it sometimes means sacrificing the depth found in specialized models.

Resource Consumption

When it comes to resources, LLMs can be resource-hungry. Training a model like ChatGPT demands thousands of GPUs and substantial cloud resources, which can be costly for many organizations.

In contrast, SLMs are much more accessible. Models like Mistral 7B can be run on local machines with decent GPUs, making them more suitable for smaller businesses or those just getting started with AI.

Bias and Representation

Bias is a crucial consideration. LLMs often reflect biases present in their training data, which can include a skewed representation of different groups and ideas. This is due to their training on vast amounts of publicly available information that may not always be balanced.

SLMs, with their focus on narrower datasets, typically face less bias risk. Their specialized training helps ensure more accurate and fair outcomes, making them a compelling choice for sensitive applications.

Making the Right Choice

Inevitable to Choose the

RIGHT GENERATIVE AI APPROACH

LLMs or **SLMs**

- Specialization & Customization
- Performance & Accuracy
- Resource Usage & Cost
- Security & Privacy
- Adaptability & Latency

When deciding between LLMs and SLMs, consider what you need for your organization. If you're looking for cutting-edge performance and versatility, LLMs are the way to go. However, if your focus is on specific tasks within a domain, SLMs can provide efficient, cost-effective solutions.

A smart strategy might be to start with LLMs for rapid development and prototyping. As you refine your needs and gather insights, you can transition to SLMs for optimization in specific areas, ensuring that your AI efforts align closely with your business goals.

Product Lifecycle Management for LLM-Based Products ↗

09

Learnings from Generative AI Implementation Failures

1. AI-Powered Diagnostic Tool

The project aimed to build an AI-powered diagnostic tool to detect early-stage lung cancer using radiology images. The AI was supposed to assist doctors by identifying nodules that could indicate malignancy, helping reduce false negatives, and improving early detection rates.

PoC Setup

- **Data:** Radiology images from a hospital database, annotated by radiologists, were used to train a deep-learning model.
- **Model:** A convolutional neural network (CNN) was used, designed to detect minute abnormalities in lung tissue.
- **Benchmark:** The system needed to achieve at least 95% accuracy in identifying malignant nodules without significantly increasing false positives.

Challenges and Failure Points

- The training dataset was limited to a specific demographic.
- Struggled in real-world clinical scenarios.
- Large amount of computational power and storage.
- Costs of scaling the AI system across multiple hospitals.

2. Customer Demand Prediction for Retail Chain

The AI solution aimed to predict customer demand for a large retail chain to optimize inventory management. The system would analyze historical sales data to forecast product demand, enabling the company to reduce overstock and stockouts.

PoC Setup

- **Data:** The model was trained on two years of sales data, including product types, regional preferences, and seasonal trends.
- **Model:** A time-series forecasting model (ARIMA + machine learning) was used to predict sales across different regions for specific product lines.
- **Benchmark:** The PoC success was defined as reducing inventory holding costs by at least 15% without impacting stock availability.

Challenges and Failure Points

- Data inconsistencies, redundancies, and missing records.
- The model struggled with predicting high-demand periods.
- Lack of Transparency in AI's decision-making process.
- The cost of deploying the AI system was deemed too high.

10

How to Make Generative AI Work for You with a Go To Market Strategy?

Successfully integrating Generative AI into your business requires a well-structured Go-To-Market (GTM) strategy. This strategy not only helps ensure the technology aligns with your organizational goals but also addresses the unique challenges of operationalizing AI models. Below are the key considerations that should be part of any GTM plan to maximize the value of Generative AI.

How to make GenAI work for you with a Go To Market Strategy

Strategic Approach & Vision to eradicate Generative AI challenges

- Importance of defining the problem and having a **Clear Understanding** of the **USE CASE.**
- **Massive Education Gap** for **Buyers** in understanding what they are purchasing.
- Build vs. Buy: Challenge in commercializing and operationalizing **GenAI.**
- **Open Source vs. Closed Source Models:** Pros and cons of each approach.
- **Importance** of system-wide considerations and **Deep Funnel Objectives.**
- **Data Privacy,** Compliance, and Security **Considerations.**

1. Define the Problem and Clear Use Case

A successful AI initiative starts with a clear understanding of the problem and specific use case. Generative AI can be highly effective, but only if its capabilities align with your business needs. Start by identifying areas where AI can improve efficiency, customer experience, or decision-making, and set measurable goals to ensure success.

2. Build vs. Buy: Challenges in Operationalizing AI

Deciding whether to build an AI model in-house or buy an existing solution is a critical choice. Building offers customization but requires significant resources and expertise, while buying is faster but may lack flexibility. Enterprises must weigh factors like budget, time-to-market, and scalability before making a decision.

3. System-wide Considerations and Deep Funnel Objectives

AI should align with your business's overall objectives and workflows. Integrate it across departments and define KPIs that span the customer journey, from marketing to operations. A holistic approach ensures that the AI enhances productivity and scales across the enterprise.

4. Education Gap for Buyers

There is a significant gap in understanding Generative AI among decision-makers. Buyers often lack clarity on what they are purchasing, leading to misaligned expectations. Bridging this gap requires educating stakeholders on the technology's capabilities and conducting thorough evaluations to ensure AI solutions meet business needs.

5. Open Source vs. Closed Source Models: Pros and Cons

Choosing between open-source and closed-source AI models depends on your priorities. Open-source models offer customization but require technical expertise. Closed-source models are easier to deploy with vendor support but limit flexibility and may be costly. Consider your business's capacity for AI management and long-term needs.

6. Data Privacy, Compliance, and Security Considerations

Generative AI models handle vast amounts of data, often including sensitive information. Ensuring data privacy, regulatory compliance, and robust security is essential. Implement data governance, adhere to regulations like GDPR or HIPAA, and protect systems from cyber threats to safeguard your AI solutions.

11

Key Insights for Maximizing Generative AI Value

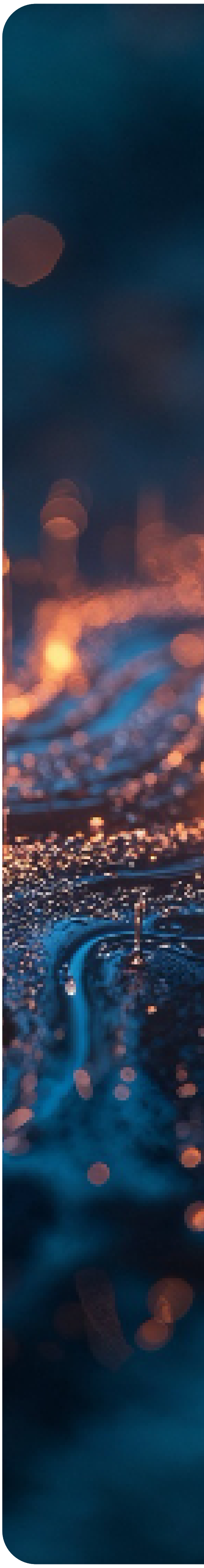
> *Two out of three (67%) say generative AI will help them get more out of other technology investments, like other AI and machine-learning models.*

Eliminate FOMO on Gen AI-Readiness

Don’t let the hype surrounding Generative AI pressure your organization into a rushed adoption. Instead, take the time to assess your current capabilities, infrastructure, and strategic objectives. FOMO can lead to wasted resources and misguided efforts. A thoughtful evaluation ensures that you’re genuinely prepared to leverage the technology effectively.

Choose the Right Generative AI Approach

Your path to Generative AI development should reflect your unique business context. Consider whether to develop AI solutions in-house or partner with established providers. Evaluate your team's expertise, data accessibility, and desired outcomes. The right approach is not one-size-fits-all; it should align with your operational needs and long-term vision.



Explore the Power of SLMs

Small Language Models (SLMs) offer a compelling alternative to their larger counterparts. While large models can be resource-intensive, SLMs deliver substantial performance with lower computational costs, making them accessible to a broader range of businesses. They excel in specific tasks like text classification, summarization, and chatbots, providing tailored solutions without the need for extensive infrastructure.

Domain-Specific LLMs in the Real World

The real-world applications of domain-specific LLMs are transformative. For instance, in healthcare, LLMs can assist in patient diagnosis and treatment recommendations by processing specialized medical literature. In finance, they can analyze market trends and offer insights tailored to financial professionals. These models not only boost efficiency but also empower organizations to make informed decisions backed by data-driven insights.

Actionable Learnings from Failure to Success Stories

The journey to Generative AI implementation is rarely straightforward. Many enterprises encounter obstacles and learn valuable lessons along the way. By studying these case studies—both failures and successes — you can glean actionable insights that inform your strategy. Understanding the pitfalls, such as overestimating model capabilities or underestimating data preparation needs, allows you to navigate challenges more effectively and increase your chances of success.

12

Final Thoughts

The adoption of Generative AI models is reshaping how enterprises operate and deliver services. Companies like Adobe and Netflix have successfully integrated Generative AI to enhance user experiences and streamline workflows.

For enterprises looking to adopt Generative AI, a collaborative approach across departments is crucial. Marketing, product development, and engineering teams should work together to identify use cases that align with business goals. By investing in the right tools and training, companies can effectively integrate Generative AI into their operations.

In summary, while the path to adopting Generative AI has its hurdles, the potential benefits are clear. Businesses that take a strategic approach to implementation can enhance their operations and better meet customer needs.

AZILEN
Engineering Excellence

Azilen Technologies is a **Product Engineering company**. We collaborate with organizations to propel their software product development journey from Idea to Implementation and all the way to product success.

From consulting to UX engineering, software design & development, test automation, DevOps, and modernization of software products, we engage with product companies to build a competitive advantage with the **right mix of technology skills, knowledge, and experience**.

Domain expertise, agile methodologies, and cross-functional teams blended in a **collaborative development approach** are our vanguards of engineering, managing, monitoring, and controlling **product lifecycles** for startups and enterprises.

Highly scalable and future-fit products that too with **faster-go-market** are what we deliver by letting in-house teams of product companies focus on **core product expansion & growth** while we manage and support the technology in parallel.

CONTACT US

Lets Connect, Collaborate and Innovate

FOLLOW US