bluealtair
Driving Digital Success

The Power & Limitations of LLMs

An Introductory Guide to Large Language Models (LLMs) for Business

By Prashant Choudhary
Sr. Manager, Data Science & AI, Blue Altair

Introduction

Long ago, humans developed spoken languages to communicate. Today, Artificial intelligence (AI) has given us language models that serve a similar purpose, providing a basis to communicate and generate new concepts.

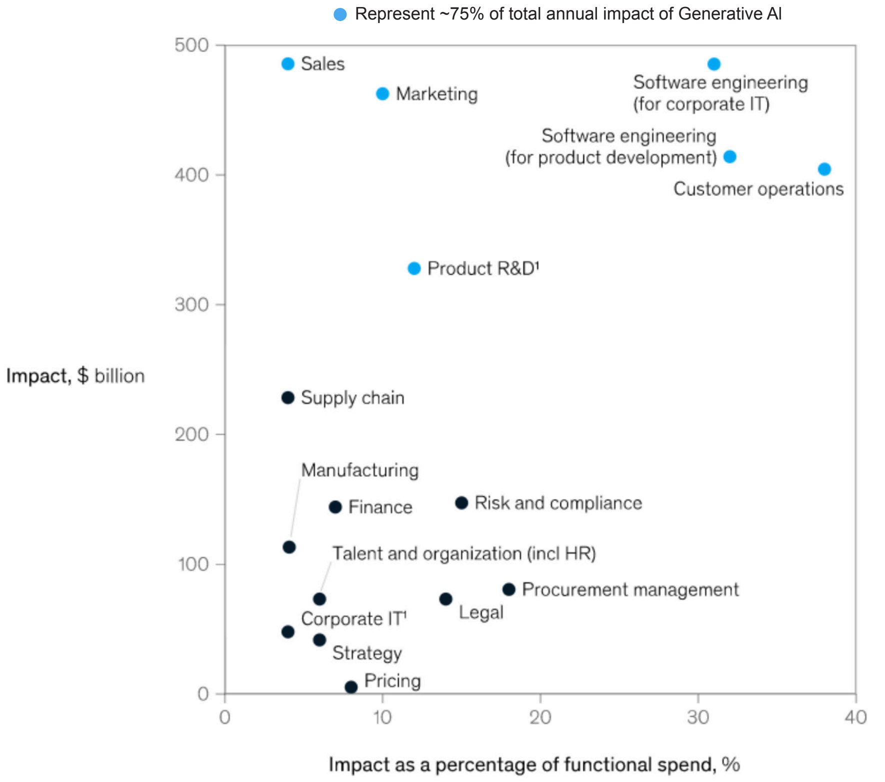
Of late, Generative AI (GenAI) has taken the world by storm, thanks to ChatGPT – one of the most exciting applications of Gen AI. Businesses across industries are paying close attention to developments in Gen AI and proceeding cautiously, given that we are in the early stages. However, with the swift pace of innovation, businesses have every reason to think big, while taking small steps now.

Based on a recent report[1] by McKinsey & Company, GenAI holds immense economic potential, with the capability to contribute between $2.6 trillion to $4.4 trillion annually to the global economy. This impact covers 63 analyzed use cases and can elevate the influence of all artificial intelligence by 15% to 40%. Additionally, the impact could double if GenAI is integrated into other software beyond the current use cases. The industries set to experience significant benefits from GenAI include banking, high tech, and life sciences. Across the banking industry, for example, the technology could deliver value equal to an additional $200 billion to $340 billion annually if the use cases were fully implemented.

To understand the possibilities that exist with Gen AI, we need to understand the source of the power behind it. Large Language Models (LLMs) have been the driving force enabling a variety of applications like ChatGPT that are and will have a significant impact on how we live and work in the coming years. However, while we are in the early days of GenAI exploration, there are many use cases already surfacing that give valuable insight into understanding both the advantages and potential pitfalls when training LLMs.

In this guide, we will focus on the maturity of LLMs and review different types that businesses can consider using, along with techniques such as prompt engineering and fine-tuning, which can be used to optimize AI performance and output. We will also touch on the best practices and industry specific use cases of LLMs.

The Economic Potential of GenAI



Note: Impact is averaged.
[1] Excluding software engineering.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

Generative AI refers to a category of artificial intelligence models that is not limited to language models but can also be used to generate images or audio.

[1] *https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/The-economic-potential-of-generative-AI-The-next-productivity-frontier?*

What are LLMs

LLMs are specific types of GenAI models that are capable of understanding and generating human-like language responses. They are massive neural network-based models that use complex algorithms to analyze and understand language patterns, and then generate responses based on that analysis. However, LLMs require a large amount of training data and computing power to learn, understand, respond, and predict with more or less the same ease as humans.

This combination is already showing promising results where businesses have automated language-based tasks. LLMs have been able to:

- Generate coherent and meaningful sentences and paragraphs of text.
- Translate between languages.
- Answer questions and provide information.
- Summarize large amounts of information.
- Write emails, compose music, or generate and debug code.

Evolution of LLMs

Researchers have done an excellent job in a paper called "Harnessing the Power of LLMs in Practice" [2] where they explain the different types of LLMs and how they have evolved.

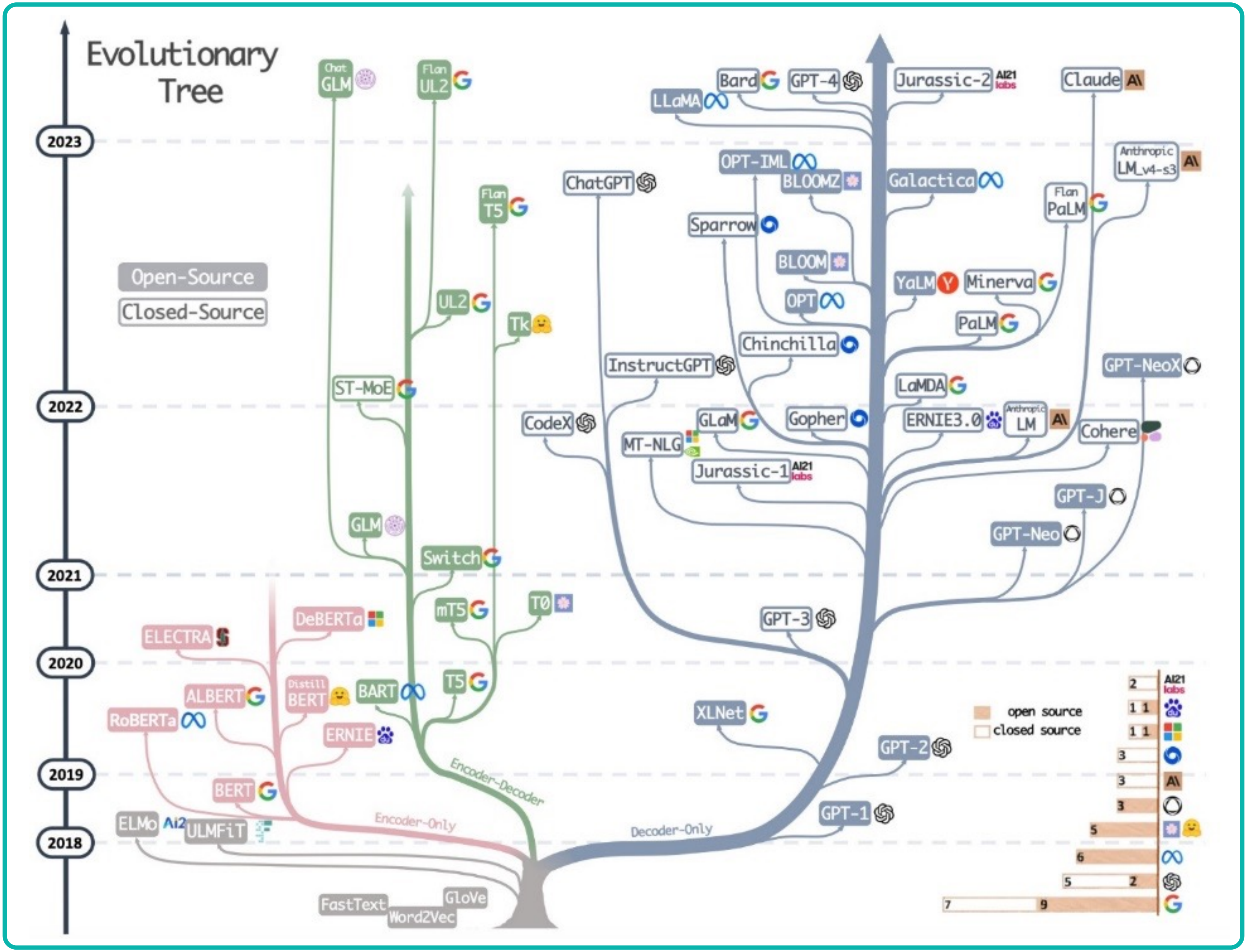
The "family tree" of modern language models shows how LLMs have evolved from models like Glove, Word2vec to advanced and powerful transformer-based models. Transformer based models can be classified into three main types: decoder-only, encoder-only, and encoder-decoder models.

Encoder-only models are well suited for tasks like text classification and entity recognition. Decoder-only models like GPT and Llama are very popular for text generation tasks and can now be generalized for most tasks.

Encoder-Decoder models can be used to model complex mappings from one sequence of text to another and are suitable for tasks like language translation and summarization.

LLM Evolutionary Tree

In the chart pictured on the following page, each branch is color-coded to represent different types of models with the newest models represented higher up on the tree. Blue represents a decoder-only model, pink stands for an encoder-only model, and green indicates the combined encoder-decoder models. The solid labels indicate an open-source model, and the hollow labels identify a closed-source model. While this chart is everchanging, the bar graph at the bottom identifies current key players in the evolution of these models.



[2] https://arxiv.org/abs/2304.13712

Commercial vs Open Source LLMs

An important decision businesses need to make is whether to embrace the power of a commercial LLM or venture into the realm of open source LLMs. However, there exists a third alternative: training your very own LLM. This option will give you unparalleled control over your LLM's performance, flexibility, and training data but it comes with a considerable price tag and inherent risks. Hence, this path is recommended only if LLMs plays a pivotal role in your business strategy and serve as a technological moat.

Commercial LLMs

Commercial LLMs like ChatGPT and Bard are ideal when you want to quickly build downstream applications and explore the vast possibilities offered by LLMs. They offer access to the best-in-class LLMs renowned for their exceptional performance, enabling you to outsource the LLM technology and focus on achieving your desired outcomes. Additionally, commercial LLMs prove valuable in scenarios where training datasets are limited, as they possess the capability for zero or few-shot learning, meaning they can make accurate predictions with minimal training examples.

Advantages

- **Simplified LLM Training:** Fine-tuning modules are simple to use for developing enterprise applications.
- **Cost-Effective:** Initial training and exploration costs are minimal, given that the major expenses occur during inference stage.
- **Data-Efficient:** LLMs require only a few examples, or sometimes none, to deliver reliable inference results.
- **Market-Leading Performance:** Helps reduce the time it takes to bring your apps to the market.
- **Enhanced App Performance:** Enhances the performance of your enterprise applications, delivering more accurate and relevant results to your B2B and/or B2C users.

Disadvantages

- **Costly Commercial Services:** If you have a high volume of fine-tuning or inference tasks, the expenses associated with commercial LLM services can be quite high.
- **Compliance Constraints:** Certain industries and use cases prohibit the use of commercial LLM services due to data sensitivity and the need to protect personally identifiable information (PII).
- **Dependency Risks:** If you rely heavily on external LLM service technology for building external apps, it’s important to consider alternative strategies to mitigate risks and establish safeguards for your business.
- **Lack of Customization:** Depending on the commercial LLM service you choose, you may have limited control over fine-tuning and customization options, which can impact the alignment of the LLM with your specific business needs.

Open Source LLMs

Open source LLMs like Llama and Vicuna offer unparalleled control over the model, granting you the freedom to customize and adapt it according to your specific needs. They are especially suitable to industries and use cases where regulatory requirements prevent the sharing of sensitive user data with commercial LLM services.

Advantages

- **Greater Control and Independence:** By opting for open source LLMs, you become less reliant on the future direction of LLM service providers. This gives you increased control over the roadmap and ensures compatibility with your existing infrastructure.
- **Community Collaboration:** Open source LLMs foster a vibrant community of developers and experts who collaborate, share insights, and collectively enhance the capabilities of the models, providing a supportive ecosystem for growth and improvement.

Disadvantages

- **Domain Expertise Requirement:** Effectively training, fine-tuning, and hosting an open source LLM demands substantial domain expertise.
- **Performance Lag:** Commercial models typically outperform open-source models by months or even years. If your competitors leverage commercial LLMs, they may have an advantage in LLM technology, prompting the need for alternative competitive advantages to differentiate your offerings.
- **Longer Time-to-Market and Reduced Agility:** Building downstream applications with open source LLMs may result in a slower time-to-market and reduced agility due to a more vertical tech stack. It requires careful consideration and planning to ensure efficient integration and seamless deployment.

LLM Fine-Tuning vs. Prompting

LLMs are highly capable systems equipped with vast knowledge. When we want to use these models for a specific task, then we must give them some details or instructions, which we call a "prompt." This prompt acts as an input or query to elicit a specific response from the model. In other words, prompting helps guide language model behavior by adding some input text specific to a task without retraining a model.

You will frequently encounter situations where the model does not produce the outcome that you want on the first try. In such situations we may have to provide examples or additional instructions in the prompt. If that doesn't work, then we may have to fine-tune the model by training the model using new data to make it more capable of the task you want it to perform. It is important not to confuse fine-tuning with prompting.

Fine-Tuning

Fine-tuning is a critical step in maximizing the performance and adaptability of LLMs for your specific needs. It allows you to mold these powerful models to better understand and generate content relevant to your domain or use case. Fine-tuning involves training your model on a set of data, so that it can learn and improve its responses without the need for explicit instructions or prompts for each individual sample. By incorporating more examples into your model's training data, you can improve its overall performance. These examples become part of the model's internal knowledge, allowing it to better understand and respond to new inputs in the future.

Advantages

- **Task-Specific Performance:** Fine-tuning allows for training the model on specific datasets, enabling it to learn task-specific patterns and improve performance on a particular domain or task.
- **Greater Control:** It gives more control over the model's behavior by directly updating its parameters. This allows for fine-grained adjustments and optimizations to align the model with the desired outputs.
- **Improved Generalization:** Fine-tuning on relevant data can help the model generalize better to new examples and perform well on specific tasks or domains.

Disadvantages

- **Resource-Intensive:** Fine-tuning a large language model like GPT-3 requires substantial computational resources, including specialized hardware and time-consuming training procedures.
- **Domain-Specific Bias:** Fine-tuning on specific datasets can introduce biases present in the training data, so care must be taken to ensure the data used for fine-tuning is representative and unbiased.
- **Lack of Flexibility:** Once a model is fine-tuned, making further modifications, or adapting it to different tasks requires retraining the model from scratch or using transfer learning techniques, which can be time-consuming and costly.

Prompting

Prompting refers to the process of providing specific instructions to your model for each sample to guide its response. Prompt engineering is used to update or try various prompts to improve the performance of a model.

Advantages

- **Flexibility:** It allows for quick and easy modifications to the model's behavior by adjusting the instructions or queries provided as prompts. This makes it convenient for users (who don't have access to the model's architecture or training pipeline) to use a model without training it.
- **Rapid Iteration:** It enables rapid experimentation and iterative improvements. Users can quickly refine and iterate on prompts to achieve desired results without going through the time-consuming fine-tuning process.
- **No Overfitting:** It does not involve updating the model's parameters or training on a specific dataset, reducing the risk of overfitting to a particular task or domain.

Disadvantages

- **Limited Control:** Prompt engineering relies on crafting effective prompts, which can be challenging and requires domain expertise. The model's responses may still be influenced by pre-existing biases or limitations in the underlying language model architecture. If you want to use a newer model, there's no way to guarantee that all your prompts will still work as intended with the newer model.
- **Lack of Generalization:** Since prompt engineering does not involve updating the model's parameters, it may not generalize well to unseen or complex tasks. The model's understanding of the task is primarily based on the input prompts, which may limit its performance on novel scenarios.
- **Unintended Side Effects:** Adjusting prompts to achieve specific outcomes can sometimes result in unintended consequences or biased outputs. Care must be taken to ensure that prompt modifications do not lead to undesirable behaviors or reinforce existing biases.

Comparison Summary

In summary, prompt engineering offers quick iterations and flexibility, but may have limited control and generalization. Fine-tuning provides more control and better task-specific performance but requires more resources and can introduce biases. The choice between the two depends on the specific use case, available resources, and desired outcomes. There are efficient methods available for fine-tuning LLMs that aim to minimize computational and resource requirements while maintaining high performance. These methods include prefix-tuning, adapters, Low-Rank Adaptation (LoRA) and recently Quantized LoRa (QLoRa). These approaches are especially beneficial for small businesses and individuals with limited budgets, as it enables the creation of robust and commercially viable LLMs.

Concerns about LLMs

As LLMs become more advanced in their ability to understand, summarize, generate, and predict new content, we are likely to see businesses across industries embrace them as an integral part of their technology ecosystem. However, we need to be aware about some the following aspects when working with LLM:

- **Biased Training Data**
  Language models can perpetuate biases and discrimination present in the training data. Overcoming this challenge requires careful data curation and implementation of bias detection and mitigation techniques.

- **Environmental Footprint**
  Training LLMs consume substantial computational resources, resulting in high energy consumption and carbon emissions. Efforts are underway to develop energy-efficient and sustainable training methods to minimize the environmental impact.

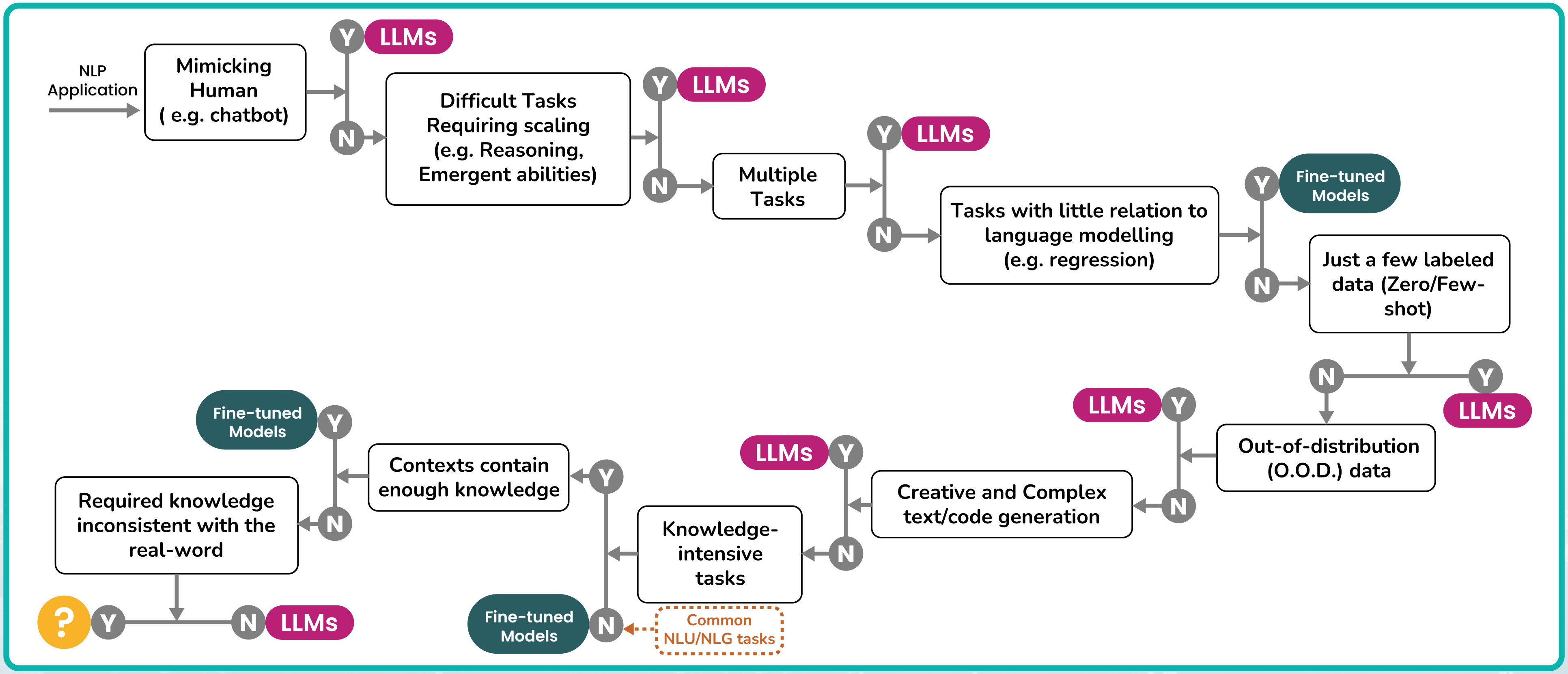
- **Lack of Transparency in Decisions**
  Despite generating impressive text, LLMs lack interpretability, making it difficult to understand their decision-making process. This raises concerns about trust, auditability, and compliance with upcoming AI regulations, especially in sensitive domains like finance and healthcare.

- **Privacy and Data Security Concerns**
  These concerns become prominent when processing sensitive information using LLMs.

- **Context Window Constraint**
  LLMs have a maximum input capacity referred to as the context window. It encompasses inputs, outputs, and additional data passed to the model, posing limitations on the amount of information the model can handle.

LLM Decision Flow

This decision flow assists users in evaluating whether their Natural Language Processing (NLP) applications fulfill specific conditions, and based on this assessment, it helps them determine the most appropriate choice between LLMs and fine-tuned models. In the illustration below which describes the decision process, "Y" represents meeting the condition, while "N" represents not meeting the condition. The yellow circle next to the "Y" in the final condition indicates that a well-performing model is not available for this type of application.


NLP Application
Mimicking Human (e.g. chatbot)
Y
LLMs
N
Difficult Tasks Requiring scaling (e.g. Reasoning, Emergent abilities)
Y
LLMs
N
Multiple Tasks
Y
LLMs
N
Tasks with little relation to language modelling (e.g. regression)
Y
Fine-tuned Models
N
Just a few labeled data (Zero/Few-shot)
N
Y
LLMs
Out-of-distribution (O.O.D.) data
LLMs
Y
N
Creative and Complex text/code generation
LLMs
Y
N
Knowledge-intensive tasks
Y
Common NLU/NLG tasks
N
Fine-tuned Models
Contexts contain enough knowledge
Fine-tuned Models
Y
N
Required knowledge inconsistent with the real-word
?
Y
N
LLMs

LLM Best Practices

- ### Explore Pre-Trained LLMs
  Use LLMs that meet your requirements otherwise consider fine-tuning it.
- ### Pick Compact LLMs
  Choose relatively smaller LLMs that meet your requirements. Most of the time, an LLM is available at several parameter sizes. For example, LLaMa is available at several 7 billion, 13 billion, and 65 billion parameters. Smaller models will help predict faster and require fewer hardware resources for training and inference.
- ### Fine-tune to Customize and Specialize
  This involves tailoring and training an LLM on your data to meet your specific needs. Using cloud-based infrastructure allows for ease of use, flexibility, and the option to pay only for what you use.
- ### Optimize Your Models for Size and Speed
  This helps ensure efficient storage and processing capabilities, facilitating faster and more responsive AI applications.
- ### Test Coverage for Reliable Results
  Test coverage refers to the breadth of test cases and prompts that encompass the range of inputs and scenarios encountered by the model. Strive for an evaluation set that accurately reflects the tasks users perform in the system, ensuring reliable and robust performance.
- ### Gather User Feedback Seamlessly
  Use low-friction methods such as thumbs up/down or short messages to gather valuable user feedback. This will help in enhancing the model's understanding of user preferences and improving its performance over time.

Use Cases for LLMs

We are beginning to see many use cases emerging for LLMs. The possibilities of GenAI are boundless. Its applications span from text and code to images, video, speech, 3D, and beyond, making a significant impact across various industries. They are being used for a variety of applications such as:

Content Generation and Marketing:

- Generating product descriptions and marketing copies
- Creating an all-encompassing AI assistant that can cater to specific user needs such as scheduling, note-taking, email management, and more

Information Processing and Comprehension:

- Summarizing research papers or legal documents for faster comprehension
- Generating reports and analysis from large datasets

Customer Support and Engagement:

- Automating customer service chatbots

Language Translation and Communication:

- Translating documents and communications for global operations

AI-Assisted Task Management:

- All-encompassing AI assistant that can cater to specific user needs such as scheduling, note-taking, email management and more

Industry Specific Utilization of LLMs

Accenture research shows 90% of all working hours in the banking industry can be impacted by LLMs, with 54% of the industry's work time having a higher potential for AI automation[3]. To fully capitalize on its potential, banks and pharma companies must embrace exploration and experimentation today to unlock greater rewards in the future.

Banking ▶ Personalized Financial Advisory

GenAI can revolutionize the way banks provide financial advisory services to their customers. By leveraging customer data, transaction history, and economic trends, AI-powered virtual advisors can generate personalized financial plans and investment strategies tailored to individual goals and risk tolerance.

How It works ▶

GenAI models analyze a customer's financial data, investment preferences, and long-term goals. Based on this information, the AI system generates personalized financial recommendations, such as investment portfolios, retirement plans, and savings strategies. The virtual advisor can engage in natural language conversations with customers, answering their queries and offering real-time advice.

Benefits ▶

- Enhanced customer engagement and satisfaction through personalized recommendations.
- Increased customer retention and loyalty as they actively participate in investments and savings.
- Potential for higher investment returns due to well-informed decisions.

Costs ▶

- Initial investment in AI technology and infrastructure, and ongoing maintenance and updates to keep the AI system effective and secure.

Estimated ROI ▶

- Improved customer retention and increased investments may lead to a revenue boost of 5-10%.

Banking ▶ AI-Powered Enterprise Search

GenAI-powered Enterprise Search optimizes information retrieval by evaluating multiple sources and summarizing results, surpassing the capabilities of current search systems.

How it Works ▶

GenAI scans and comprehends diverse data sources, including documents, databases, and web pages. Natural language processing enables the AI system to understand context and relevance effectively and generates concise and relevant summaries of information, providing actionable insights.

Benefits ▶

- Faster and more accurate information retrieval, outperforming traditional search systems.
- Comprehensive and summarized results enable quicker decision-making for employees and customers.
- Improved data accessibility enhances operational efficiency and customer service.

Costs ▶

- Initial investment in AI integration and infrastructure, and ongoing maintenance and periodic updates to ensure optimal performance.

Estimated ROI ▶

- Estimated ROI of 15-20% through increased productivity and better decision-making.

[3] *https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf#zoom=40*

Life Sciences ▸ Automating Preliminary Drug Discovery Screening

Automating Preliminary Drug Discovery Screening using GenAI accelerates the drug discovery process in the life sciences industry. This process efficiently screens potential drug compounds, leading to faster identification of promising candidates for further research.

How It Works ▸

GenAI models analyze vast datasets of chemical structures, biological interactions, and known drug compounds. The AI system identifies potential drug candidates based on their structural properties and predicted interactions with target molecules. These AI-generated drug candidates undergo initial screening, enabling researchers to focus on the most promising candidates for further testing.

Benefits ▸

- Improved Speed: GenAI expedites the drug discovery process, reducing the time required to identify potential candidates.
- Enhanced Quality: By analyzing vast amounts of data, AI improves the accuracy and reliability of preliminary screening.
- Resource Efficiency: Automation reduces the need for labor-intensive manual screening, saving time and resources.

Costs ▸

- Implementing and integrating GenAI for drug screening may require initial investments in AI infrastructure, software development, and data management.

Estimated ROI ▸

- The use of GenAI in automating preliminary drug discovery screening can yield an estimated ROI of 2.6% to 4.5% of annual revenues in the pharmaceutical and medical-product industries.

Life Sciences ▸ Patient Engagement and Virtual Assistants

Patient Engagement and Virtual Assistants powered by LLM technology transform healthcare interactions by offering personalized health information and guidance to individuals. LLM-based virtual assistants understand natural language queries and medical data, enabling personalized recommendations, medication adherence support, and lifestyle advice, leading to increased patient engagement.

How It Works ▸

LLMs analyze patients' medical history, symptoms, and preferences to comprehend natural language queries. Based on this information, the virtual assistant provides personalized health recommendations and lifestyle advice. The virtual assistant continuously learns from interactions, improving its ability to cater to individual patient needs.

Benefits ▸

- Personalized Care: LLM-based virtual assistants offer tailored health guidance, improving patient satisfaction.
- Improved Medication Adherence: Reminders and support from virtual assistants enhance medication adherence.
- Reduced Manual Interventions: Automation reduces the workload of healthcare providers, streamlining operations.

Costs ▸

- Initial costs include the development and integration of LLM-based virtual assistants into existing healthcare systems, and ongoing expenses may include maintenance and updates.

Estimated ROI ▸

- The implementation of Patient Engagement and Virtual Assistants using LLM technology can yield a considerable ROI due to increased patient satisfaction, better health outcomes, and optimized healthcare workflows.

In Conclusion

LLMs have undeniably revolutionized the landscape of NLP. Their versatility spans a multitude of use cases, from generating creative content to aiding in complex problem-solving. While they are powerful tools, it is essential to acknowledge their current shortcomings and take appropriate measures to ensure their ethical use. With responsible usage and continual research and development, we can harness the power of LLMs to drive innovation, improve communication, and find solutions to complex problems.

The Blue Altair Advantage

With a robust track record of more than 7 years, Blue Altair stands out as a leader in NLP, with a significant focus dedicated to this space. Our expertise spans various areas encompassing language models, named entity recognition, classification, summarization, and Q&A chatbots. Beyond development, we have successfully deployed models in production. Our commitment extends beyond intelligent language model use, ensuring content safety for end-users.

Notably, Blue Labs, our research and innovation wing, has undertaken various projects, some of which have accelerated client use cases. We have executed several projects using both open source and Propriety LLM models. As the R&D division of Blue Altair, Blue Labs makes sure that we are up to date on the latest and greatest technological advancements across our capabilities.

Most of our clients have been associated with us since our inception and consider us an extension of their team. Our long-lasting client partnerships are based on our ability to adapt to their ever-changing needs and a strong commitment to their success. Our accomplishments have been recognized in the industry, earning us inclusion in IDC Innovators in Artificial Intelligence Services in 2020 and Forrester Now Tech™ for AI Consultancies in 2021.

About the Author

Prashant brings over 8 years of experience in developing innovative, data-driven solutions for complex business problems. Currently, he works as a Sr. Manager – Data Science & AI for Blue Altair. Prashant's expertise includes all aspects of advanced analytics including descriptive, predictive, and prescriptive analytics, along with proficiency in machine learning, deep learning, knowledge graphs, natural language processing, large language models, and big data.

Prashant has successfully built data-driven solutions across diverse service sectors with a major focus in the life sciences domain.

bluealtair

Driving Digital Success

United States | India | Australia | Argentina

www.bluealtair.com