FUJITSU

Leveraging the LLM: Strategy from Model Selection to Optimization

Insight for top management

Contents

Leveraging the LLM:
Strategy from Model Selection to Optimization
–Insight for top management–

Leveraging the LLM: Strategy from Model Selection to Optimization
–Insight for top management–

1. Introduction: the Continually Expanding Potential of LLMs

The arrival of ChatGPT or Google Gemini together with its rapid adoption marked a key turning point in the recognition of the potential of Generative AI, both transform organizations and the productivity of their employees. Business leaders are realizing that generative AI powered by large language models (LLMs) has the potential to radically transform everything from business to industry to society at large, opening new frontiers of value creation. They recognize the importance of actively participating in the value creation process through the deployment and adoption of successful use cases.

Emerging research studies are beginning to shed light on the potential economic benefits and productivity enhancements that could be realized through the adoption of Generative AI. Specifically, estimates suggest that the economic value of generative AI could reach up to $7.9 trillion per year if it were to spread across all industries.*1 In addition, a large study of GitHub Copilot users found that 1) user productivity increased by approximately 30% with nearly 30% acceptance of code suggestions, and 2) the impact on the global economy reached $1.5 trillion based on the results of this empirical study.*2 In addition, by analyzing usage data from more than 5,000 users of generative AI interaction tools, researchers showed that the use of these tools not only improves productivity by an average of 14% in terms of problems solved per hour, but also improves the customer experience and increases employee retention and motivation to learn.*3 FrameDiff, developed by MIT CSAIL, is a computational tool that uses generative AI to create new protein structures to accelerate drug discovery and improve gene therapy.*4 Moving beyond worker productivity, the impact of generative AI extends to scientific exploration research as well.

The potential power of such generative AI comes from the underlying model (this paper is limited to LLMs built on the Transformer architecture), which can outperform conventional AI. LLM technology is rapidly evolving, with improvements in model performance (such as improved accuracy) and enhancements (such as diversification of tasks that can be addressed and expansion of plug-in capabilities) leading to increased generation capabilities and business model expansion. In addition, generative AI will enhance humanity, increase accessibility, and promote "democratization," allowing new people who have not previously jumped on the digital bandwagon to join the digital transformation train. We're just at the start of the generative AI revolution, and the ceiling for economic impact, business impact, and driving innovation may be higher than you think.

Of course, it's up to the end users, such as value-driven companies, to maximize and embody the vast untapped potential and new value-creating capabilities of generative AI with LLM at its core. Let's discover how to get the most out of LLM from an enterprise user perspective.

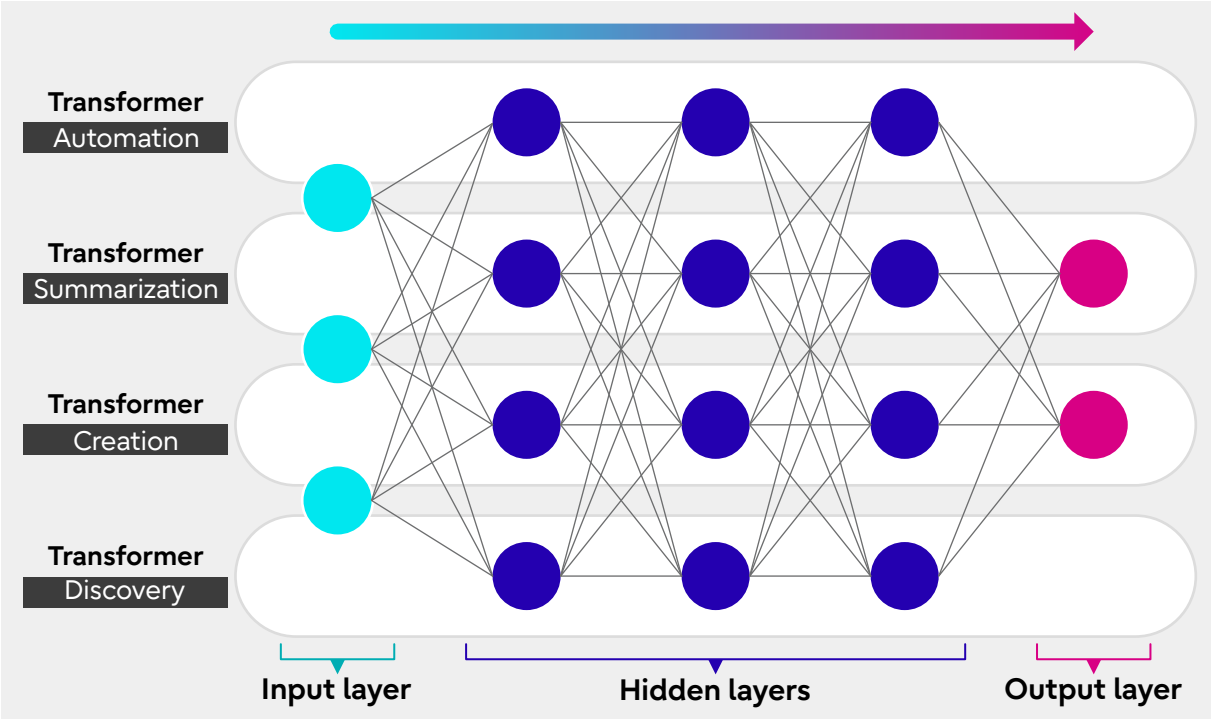
*1 McKinsey (June 2023) "The economic potential of generative AI: The next productivity frontier"
*2 Thomas Dohmke, et al. (June 2023) "Sea Change in Software Development: Economic and Productivity Analysis of the AI-Powered Developer Lifecycle"
*3 Erik Brynjolfsson, et al. (November 2023) "Generative AI at Work"
*4 MIT (July 2023) "Generative AI imagines new protein structures"

2. Uncovering the Evolving Landscape of an LLM with Superior Abilities

Conventional AI was limited to individual tasks and functions and required the development of individual models. In other words, it was primarily a point-solution approach designed to solve a single problem. Conventional AI that performs only these specific tasks, called "specialized AI" (narrow AI), uses "neural networks" or "deep neural networks",[*5] but the models have only hundreds to millions of parameters.

LLM, on the other hand, is a generic model (see Figure 1) that can handle a variety of tasks, trained on a large dataset using a new deep learning model, the Transformer. The number of parameters in the model ranges from billions to hundreds of billions. The versatility of LLM is likely due not only to the versatility of the training dataset, but also to the emergent nature of the model scale beyond a certain scale (threshold).[*6]

Figure 1 Conceptual diagram of LLM architecture
–Transformer model Neural network–



Source: Created by the author

*5 A "neural network" or "deep neural network" is a type of deep learning that allows a computer to replicate the way the neurons in a person's brain work and derive highly accurate answers.

*6 Jason Wei, et al. (2022) "Emergent Abilities of Large Language Models"

Ability of LLM

Although the specific descriptions are not unified, we summarize them in four ways considering the level of human intellectual tasks:

(1) Automation: This is the most basic level and includes data collection, organization, and simple calculations. However, with the advancement of generative AI model technology, the level of automation is progressing from Conventional AI Robotic Process Automation (RPA) to Enhanced Process Automation (EPA) to Cognitive Automation (CA).

(2) Summary: Next, you need the ability to condense large amounts of information and extract the key points. You need to understand and recontextualize information. This capability has the potential to fundamentally change the way people access and use data, from searching to questioning.*7 We believe that the "code red" that Google began to shout applies to all companies.

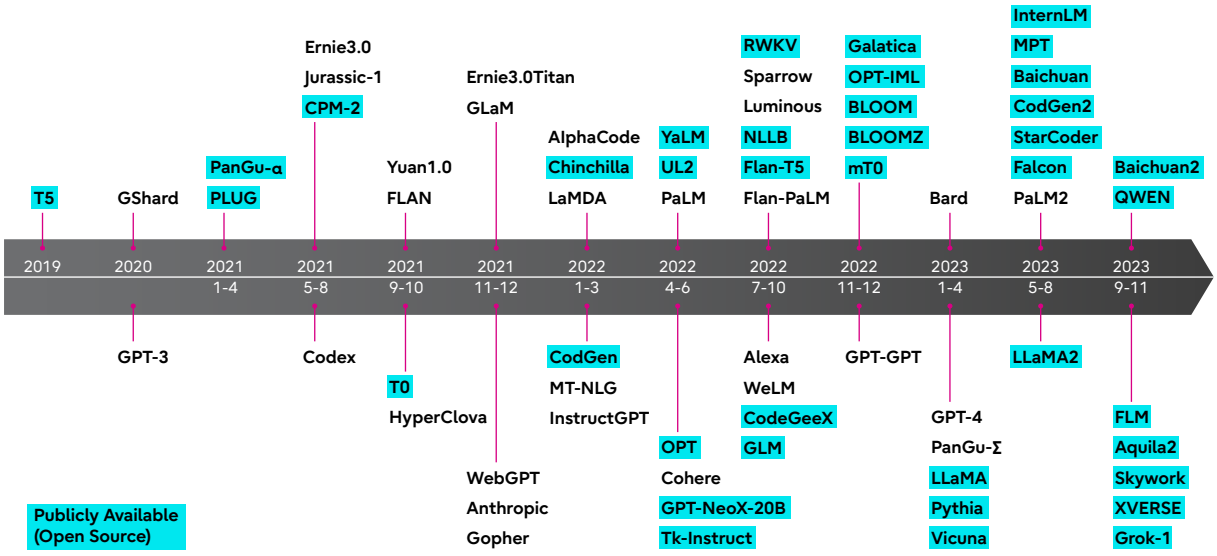
(3) Creativity: A more sophisticated level involves the ability to generate new ideas and concepts. This includes:
You must have the creativity and insight to create new information from existing data. This is such a powerful skill that it is rated as "For the first time in history, we have technology that can directly augment human knowledge creation".*8

(4) Exploration: The highest level involves the ability to discover new patterns and correlations in existing data that require deep understanding and analysis. This capability is the potential to navigate the unknown to explore the future.

The Current Landscape of Large Language Models (LLMs)

Large Language Models (LLMs) are evolving at a pace never seen in emerging technologies.

Figure 2 Existing large language models' landscape (having a size larger than 10B)



Note: Illustrations exclude company logos and names; only model names are shown.
Source: Researched and created by the author based on "A Survey of Large Language Models"

*7 In other words, by querying generative AI instead of search tools like Google, people will change their behavior.
*8 KPMG (2023) "Generative AI: From buzz to business value"

The vendor landscape for LLM development spans the academic sector, including academically based universities and research institutions, resource-rich technology giants, and innovative startups looking to push the boundaries of technology. Figure 2 shows LLMs with 10 billion or more parameters and published results, ranked by publication date of the technical paper.[*9] To put it in perspective, including other studies,[*10] there are some clear trends in the LLM landscape.

1) OpenAI and Google have established themselves as leaders in LLM development, releasing LLMs that evolve in performance over time and LLMs with new features. OpenAI extends the capabilities of its models by introducing plug-ins to ChatGPT that allow them to access the Internet, retrieve current information, perform computations, and use third-party services.[*11]

2) In the early days of LLM development, most major technology companies participated in LLM development and released models. However, since mid-2022, startup activity has increased and LLMs have played a leading role in the number of releases.

3) The models of OpenAI and large technology companies (except Meta) are essentially closed source, licensing models provided through APIs. On the other hand, Meta and many startup models are open source with local deployment. Today, more and more large technology companies are open-source models with relatively few parameters (billions to tens of billions).

4) Early in the development of LLMs, it was believed that model performance would increase proportionally to the number of parameters, and the goal was to achieve better performance by increasing the number of parameters. However, models released since 2022 have a variable number of parameters. This proved that even relatively small models (Small Language Models, or SLMs)[*12] can perform as well as larger models by increasing the training data.[*13] In other words, we have a better understanding of the scaling law, which states that there is an optimal model size (number of parameters) under given conditions (data size, computational budget, inference latency requirements, etc.). There are also empirical studies showing that SLMs can beat large models if the quality of the training data is good.[*14]

Key Challenges and Prospects for LLMs

The LLM has unprecedented potential, but at the same time it faces unprecedented or more complex challenges. The key issues are outlined below.

1) Hunger for data (Data dependency)

LLMs need a lot of data to learn and evolve. However, due to resistance to indiscriminate data collection and increased data management, LLM developers face the problem of data scarcity. In addition, the availability of computing resources and energy consumption are emerging issues.

2) Difficulty of interpretability

It is difficult to determine why or how LLM produces a certain output, called the "black box" problem, which causes social anxiety.

*9 Wayne Xin Zhao, et al. (November 2023) "A Survey of Large Language Models"
*10 Veysel Kocaman (July 2023) "Beyond OpenAI in Commercial LLM Landscape"
*11 ChatGPT plugins
*12 There is no single definition for SLM. It refers to small models (SLMs) with a number of parameters ranging from billions to tens of billions.
*13 Zian Wang (December 2023) "The Underdog Revolution: How Smaller Language Models Can Outperform LLMs"
*14 Gennaro S. Rodrigues (August 2023) "Not-So-Large Language Models: Good Data Overthrows the Goliath"

3) "Hallucination" problem
The model may produce inaccurate or misleading output. Unintentional misinformation output can cause "hallucinations" when the model produces untrue output or attempts to fill gaps in missing data. LLMs lack a built-in fact-checker, which needs to be addressed.

4) Intellectual property issues
The problem arises when AI models generate content that may infringe existing copyrights or plagiarize existing works. The issue of copyright protection is not unique to generative AI. Addressing this issue requires a combination of policy development and legal intervention from the perspective of balancing the promotion of innovation and the protection of intellectual property.

5) Toxicity issues (harmful or discriminatory output)
Toxicity in the context of AI refers to harmful or discriminatory output against specific groups of people, especially marginalized or protected groups. While toxicity issues exist in Conventional AI, LLM in generative AI can exacerbate the problem.

Various initiatives have been undertaken and progress has been made in some areas to address these issues. For example, the following approach is used to solve the "data dependency" problem.

Data augmentation:
Manipulate existing data to create new data. Image data can be rotated, scaled, inverted, and so on. This technology has reached a certain level of maturity.

Transfer learning:
Use models trained on large amounts of data and adapt them to specific tasks. This eliminates the need to collect large amounts of new data. This is an approach that leverages the capabilities of generative AI and is promising for downstream applications.

Semi-supervised learning:
This is useful when there is less labeled data (teacher data). Train the model on both labeled and unlabeled data.

Generate model data:
Once trained, LLMs can generate inexhaustible synthetic data. This allows you to increase the training data for your model.

Model-generated synthetic data is artificial data that mimics real data and can be used to train machine learning models while avoiding privacy issues. It is particularly useful for data sets that contain sensitive information or lack data that meets certain conditions or requirements. However, synthetic data also has bias issues inherited from the original data, data quality issues that do not fully mimic the real data, and overlearning issues that over-adapt to certain patterns and scenarios, so appropriate countermeasures are required.

Automated fact-checking is considered an effective approach to fact-checking and "hallucination" problems. The model determines in real time whether what you are reading is accurate. These include RAG (Retrieval-Augmented Generation), which was developed to bridge the gap between the vast knowledge of general-purpose language models and the need for accurate, contextually up-to-date information, and WebGPT, which makes generated answers more accurate and reliable by incorporating citations into the answers. Both RAG and WebGPT technologies have proven to be effective in fact checking and reducing hallucinations. The high-precision hallucination detection technology developed by Fujitsu is expected to realize reliable interactive generation AI applicable to corporate operations.[*15]

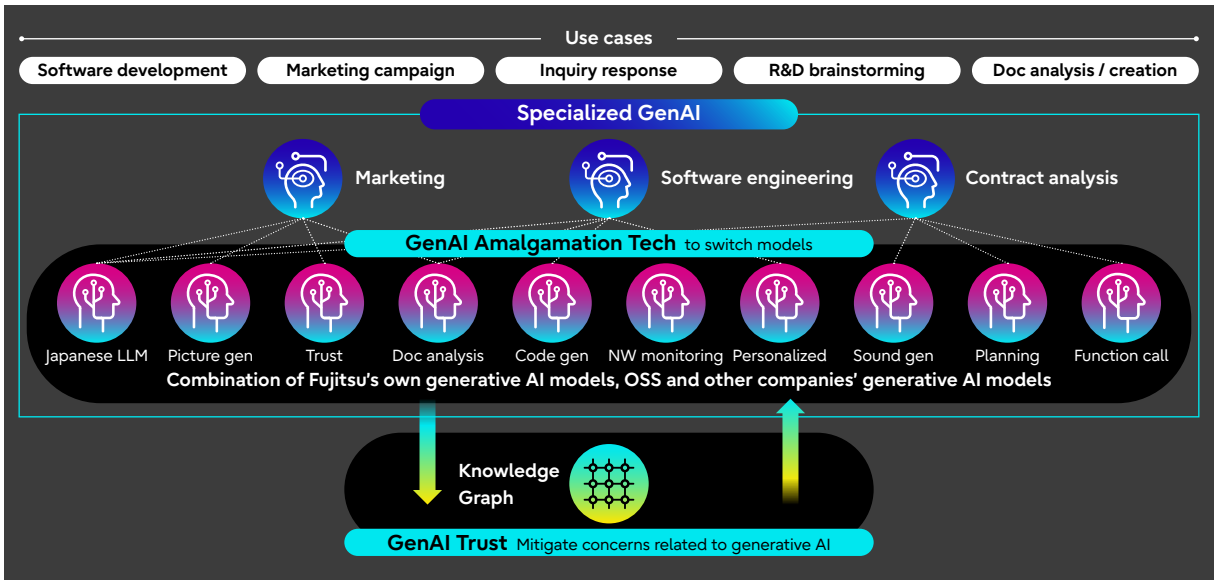
*15 Fujitsu White Paper (December 2023) "AI Trust and Fujitsu's AI Trust Technologies in Conversational Generative AI"

In addition, when considering model efficiency from the user's point of view, attention has been focused on the development of the Sparse Expert Model, which reduces the computational burden by using a subset of parameters that best fit a particular query, rather than using all parameters in the LLM. Such an expert model could either be a singular entity with multiple areas of expertise, or it could be a composite of expert models (Mixture Modal), constructed by integrating multiple SLMs, each representing a different area of expertise, tailored for a specific task. Most models with large numbers of parameters, such as Google's GLaM (1.2 trillion parameters), are considered expert models, and the introduction of expert model technology is expected to improve model efficiency and alleviate interpretability problems.[*16] In fact, the development of expert models, along with synthetic data and fact checking, is considered a major trend in LLMs.

Fujitsu has developed Generative AI Model Amalgamation Technology (Gen. AI Amalgamation Technology) as a technology that combines multiple high-performance, small-scale, specialized generative AI models (SLMs) to achieve the highest performance for use cases (specific tasks) (see Figure 3).[*17] The technology can switch between SLMs to find the optimal model in response to individual prompt inputs. In addition, the amalgamation technology has confirmed accuracy equal to or better than generative AI models including GPT-4V and HuggingGPT in image recognition tasks. In December 2023, Fujitsu developed a specialized model based on open LLM with additional learning and tuning for Japanese language performance. This Japanese model has achieved the highest level of accuracy in JGLUE evaluations.

Such progress in technological development and practical accumulation will surely advance the practical use of expert models (MoEs).

Figure 3 Conceptual diagram of Generative AI Amalgamation Technology



Source: Fujitsu

*16 Rob Toews (February 2023) "The Next Generation Of Large Language Models"
*17 Fujitsu Press Release (February 14, 2024)

3. Selection and Optimization Strategies for Leveraging LLMs

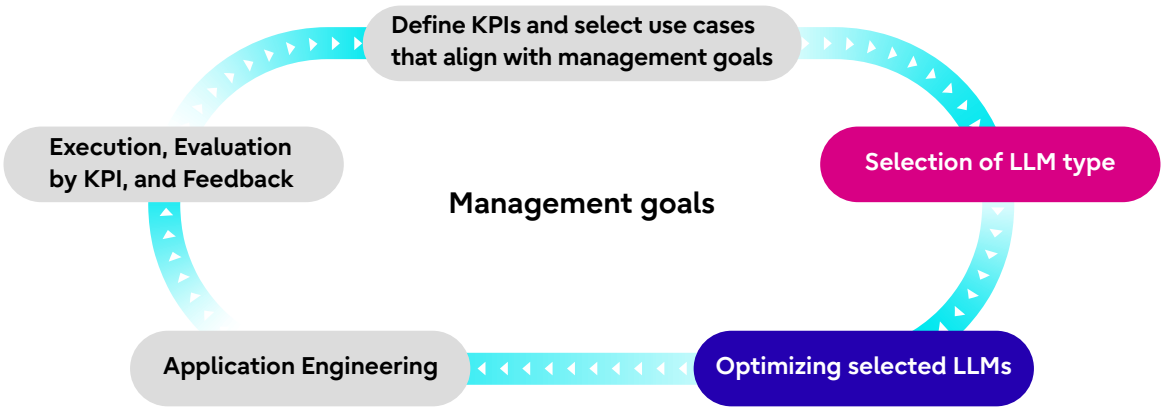
As we have seen, the LLM ecosystem is rapidly expanding and deepening. As industry leaders understand LLMs, they are accelerating their efforts to integrate the technology into their operations and usher in a new era of value creation.[*18]

LLM Lifecycle

For enterprises, LLMs are positioned as a new foundation and platform for digital transformation, not just a new toy. There are several specific and important steps in the process of creating value with LLMs. These include selecting a core model (existing or homegrown), optimizing the model, and developing use cases and applications. From a management perspective, all these processes should be guided by the value-based goals of the organization. Therefore, the first step in starting the process is to set a value objective for the entire organization.

*18 Fujitsu Blog (August 2023) "Generative AI"

Figure 4 LLM Lifecycle



Source: Created by the author

Our previous insight paper*19 focused on use cases in the value chain that leverage LLMs. However, as Figure 4 shows, this article focuses on two important steps: model selection and optimization.

Key LLM Selection Issues from an Enterprise User's Perspective

Our previous Insight paper outlined three ways to leverage LLM capabilities in terms of business goals, technical capabilities, and model diversity. These include 1) adopting off-the-shelf models, 2) integrating off-the-shelf models into our data and systems for customization, and 3) developing models that are appropriate for our purposes. Today, most companies use either 1) or 2) or a mixture of the two. However, large technology companies with large data assets and some of the largest industrial companies are also developing their own, and this trend may become more widespread as LLM technology matures and advances.

And if you look at the LLM landscape, there's a lot of activity in the development of open-source models. As of last year, there were more than 8,000 open-source generative AI projects (including models and applications) on GitHub.*20 Many working open-source models are available through hubs such as Hugging Face and GitHub. As Figure 2 shows, several LLMs with 10 billion or more parameters and published evaluation results have been released. BloombergGPT (a closed LLM with 50 billion parameters) is a prime example of LLM development for corporate financial information processing based on the open-source large language model Bloom.*21

In addition, SLMs are on the rise. In general, the larger the model size (number of parameters), the better the performance. However, it is important to understand that LLM performance is a function of several factors, including data quality, computational usage (training), and model architecture. Model performance goals can be achieved by improving other factors rather than by increasing model size. Therefore, LLM selection should also consider the cost of running the model and the overall latency rate. When considering energy efficiency and deployment on edge devices, it is important to explore models that balance performance and efficiency. In other words, the use of SLMs should be well considered.

*19 Fujitsu Insight Paper (January 2024) "Generative AI: Use Cases as the Pathway to Value Creation"
*20 Gwen Davis (October 2023) "A developer's guide to open source LLMs and generative AI"
*21 Shijie Wu, et al. (December 2023) "BloombergGPT: A Large Language Model for Finance"

Table 1 summarizes the three options for selecting an LLM. It will help you choose a model that fits your organization's purpose.

Table 1 Overview of the three LLM options

	Option 1 Use commercial LLM (via API)	Option 2 Leverage existing open LLM	Option 3 Proprietary development of LLM
Pros	• Minimize required development technologies • Save on development costs • No pre-training data required • High-quality, proven, high-level LLM available quickly • Simplifies LLM maintenance and upgrades	• LLM leverages what it has learned from massive amounts of data and doesn't have to pay for IP when making inferences. • Control over the model versus Option 1 • Reduces development time, data requirements, and training budget compared to Option 3, which builds models from scratch. • More effective for edge cases and specific applications	• Provides maximum control and flexibility over LLM performance and upgrades compared to Options 1 and 2 • Have full control over the training dataset, which directly affects model quality, bias, and toxicity issues compared to Options 1 and 2 • Protect against your own data breaches and gain a competitive advantage by training on your own data
Cons	• The total cost of ownership (TCO) making commercial LLM services with large amounts of fine-tuning and inference tasks expensive. • If you rely heavily on external LLM service, you need to find other competitive advantages when building external applications based on that LLM. At the same time, you need to reduce business risks. • Less flexibility - no support for edge inference and limited ability to customize and continuously improve models. • It prohibits the use of commercial LLM services as a use case for many industries because it is not recognized as a compliance service with sensitive data and personal information.	• While not enough to build your own, training, fine-tuning, and hosting an open-source LLM requires the skills of many domain experts. There is also a problem of LLM reproducibility. • If you're building downstream apps, the more vertical technology stack slows time-to-market and makes you less agile. • The performance of the open model typically lags behind that of the closed commercial model by several months/year, which may be a competitive disadvantage over competitors introducing Option 1.	• It is expensive, requires software, hardware, and cross-domain knowledge, and is risky. • Compared to Option 2, which can learn from valuable data on the Internet and provide a solid starting point, Option 3 starts from scratch and takes time to acquire generalized capabilities.
Selection condition	• It's a great choice if you have limited technical resources, very limited training data, and want to use high-performance LLM to build downstream apps. • It is suitable for prototyping applications and exploring LLM possibilities.	• It is better to keep the architecture of the open source model unchanged and either take the existing pre-trained LLM directly and fine-tune it, or take the weights (Parameters) of the existing pre-trained LLM as a starting point and continue the pre-training. Especially when the training data set is not large or diverse, you can take advantage of the knowledge learned by the original model. • Operating in a regulatory environment, having user/sensitive data that cannot be provided to commercial LLM services, or needing to deploy the model at the edge due to latency or location are also examples of Option 2 choices.	• If you need your own LLM architecture or want to pre-train on your own dataset, this is the best choice. • In this case, the proprietary LLM becomes a core part of the business strategy and technological competitive advantage (moat), and is the choice of companies that want to have the technological innovation and invest significantly in model development. • It is a good option if a company wants to use a lot of proprietary data to run a continuous model improvement loop for sustainable competitive advantage.

Source: Created with author additions based on Rebecca Li, et al. (2023) "Current Best Practices for Training LLMs from Scratch", Gwen Davis (October 2023) "A developer's guide to open source LLMs and generative AI".

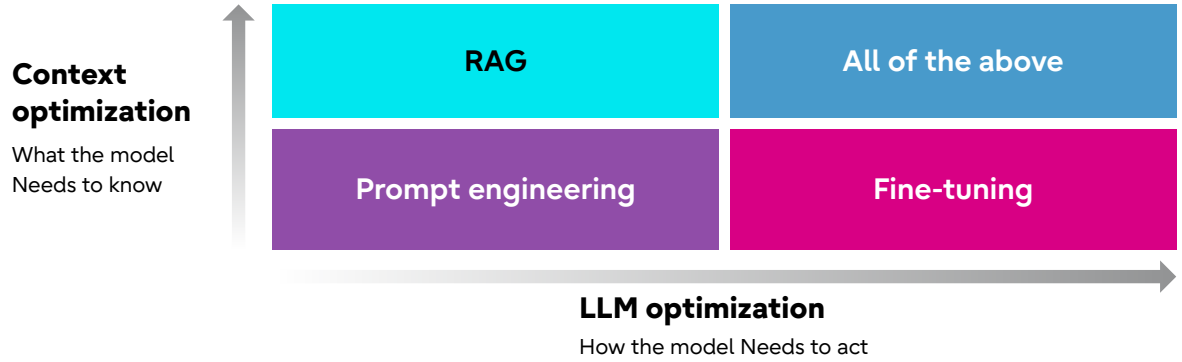
Optimizing Selected LLMs

Pre-trained on large datasets, LLMs can take advantage of transfer learning and emergent capabilities to solve many tasks. However, for enterprise users, optimization is the next step in LLM selection, not only to adapt LLMs to their new use cases, but also to demand higher performance (e.g., accuracy) and from a compliance perspective.

Broadly Defined LLM Optimization

As shown in Figure 5, LLM optimization is broadly defined. Basically, there are two main methods: 1) "what the model should know", i.e., contextual optimization by acquiring knowledge other than pre-trained information, and 2) "how the model should behave", i.e., LLM optimization in the narrow sense by adjusting the behavior of the model. Of course, a combination of 1) and 2) is also possible. The method of LLM optimization is chosen based on requirements such as the structure of the model and the company's own goals and resources.

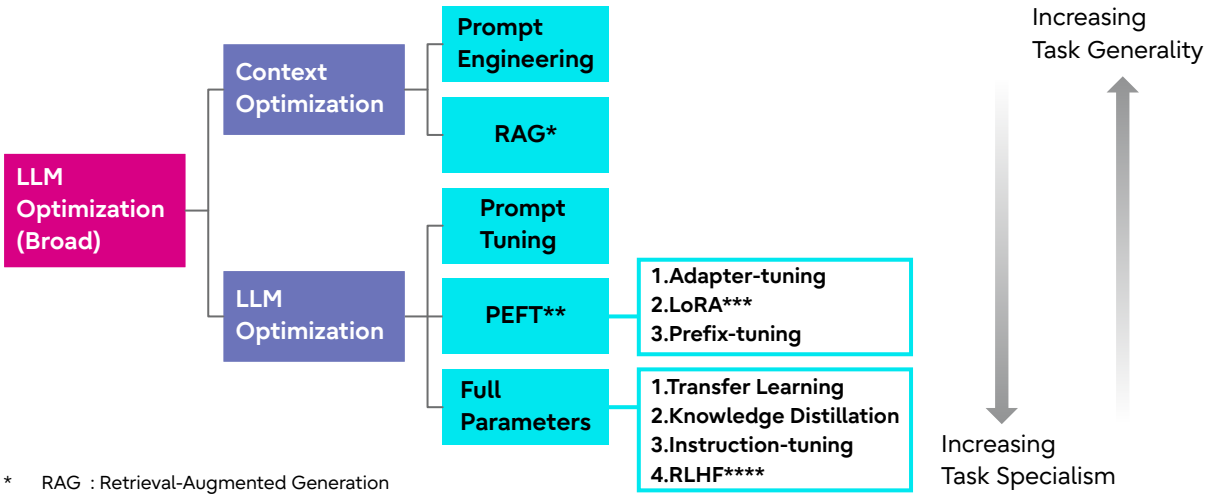
Figure 5 Two methods and mix method of LLM optimization



Source: Open AI (November 2023) "A Survey of Techniques for Maximizing LLM Performance"

A broad overview of the major LLM optimization methods and techniques is shown in Figure 6.

Figure 6 Main approach of LLM optimization (Broad)



* RAG : Retrieval-Augmented Generation
** PEFT : Parameter-Efficient Fine-Tuning
*** LoRA : Low-Rank Adaptation
**** RLHF : Reinforcement Learning from Human Feedback

Source: Created by the author

What is Contextual Optimization?

Contextual optimization in LLMs is a technique that optimizes the performance of a model by adjusting the input data (or "context") without making direct changes to the model itself. This is especially useful when commercial LLMs have limited direct access to the model, or when it is difficult or impossible to retrain the model.

Contextual optimization basically consists of 1) "prompt engineering", which improves the output of the model by adjusting the input prompts to the model, and 2) techniques that bridge the gap between purely generative models and external knowledge by capturing knowledge from a variety of outside sources, such as databases and articles, and bringing the necessary information into the model to improve contextual relevance and factual accuracy. 2) is commonly referred to as RAG (Search Extended Generation). RAG is particularly useful in cases where real-time data access, domain expertise, or fact checking is required. This is especially useful for tasks where the model requires up-to-date information, or for use cases where strong anti-hallucination measures are required.

The approach of optimizing an existing LLM through RAG, which learns from such direct input and previous interactions (contextual learning) and retrieves the most relevant information from a large dataset, requires much less time, computational resources, and even less advanced expertise than training an LLM from scratch or fine-tuning the model we are about to describe (LLM optimization in the narrow sense). Big tech companies like Microsoft, Salesforce, and Tesla are already using this approach.

LLM (Narrowly Defined) Optimization

Unlike contextual optimization, LLM (narrowly defined) is a direct modification of the model itself. To achieve greater accuracy in each domain, or when contextual optimization does not achieve the desired goal, LLM must be adapted to downstream tasks through tuning, which modifies the parameters of the model. This is because the LLM parameters determine the overall behavior of the model, and changing the parameters also changes the behavior of the model.

Figure 6 shows three commonly used methods for LLM optimization (narrowly defined). Depending on how much you change the model itself, it can be classified as 1) prompt tuning, 2) parameter efficient fine tuning (PEFT or partial tuning), or 3) full fine tuning. Without going into technical details, here are the main methods:

1) Prompt tuning differs from prompt engineering in that it introduces additional parameters into the prompt and optimizes those parameters in a supervised sample.

2) PEFT is a technique that improves model performance by changing only some of the model parameters. The primary methods are adapter tuning, LoRA, and prefix tuning.

3) Full fine-tuning updates all model parameters. Key methods include transfer learning, knowledge extraction, instruction tuning, and reinforcement learning with human feedback.

Overall, narrow-sense LLM optimization, which modifies LLM parameters, is a good approach to model tuning when the model is small and there are few tasks involved. However, full fine-tuning with a relatively small dataset can lead to "disruptive forgetting",[*22] or a model that excels at certain tasks but becomes less proficient at others. If you need the ability to perform multiple tasks simultaneously, multitask learning should be used, and full fine-tuning of single tasks should be avoided. On the other hand, if the size of the pre-learned LLM is large enough, the context-optimized tuning approach may yield comparable performance, since the underlying behavior of the neural network is to activate the features and information (required parameters) needed to perform a particular task.[*23] These results have important implications from an enterprise management perspective.

LLM Performance Evaluation

LLM Performance Evaluation is repeated as you optimize the selected LLM until you reach your goal. The performance of the model is evaluated on a test set based on predefined KPIs.

The performance that an organization can directly address is not only related to the model, but also to the selection of use cases and the development of applications that meet management objectives. Therefore, we have already published an insight paper summarizing the KPIs.[*24] There are eight KPIs that are closely related to model performance: 1) quality KPIs (turnaround time, accuracy, output quality and error rate) and system KPIs (training time and cost, human participation metrics, scalability, corporate compliance risk management). Of course, it is difficult to achieve all these indicators at the same level, and we believe that we should prioritize and evaluate them in line with your management objectives.

*22 Everton L. Aleixo, et al. (2023) "Catastrophic Forgetting in Deep Learning: A Comprehensive Taxonomy"
*23 Weights & Biases (2023) "Best Practices for Fine-Tuning and Prompt Engineering LLMs"
*24 See footnote 15.

4. Recommendations for Top Management

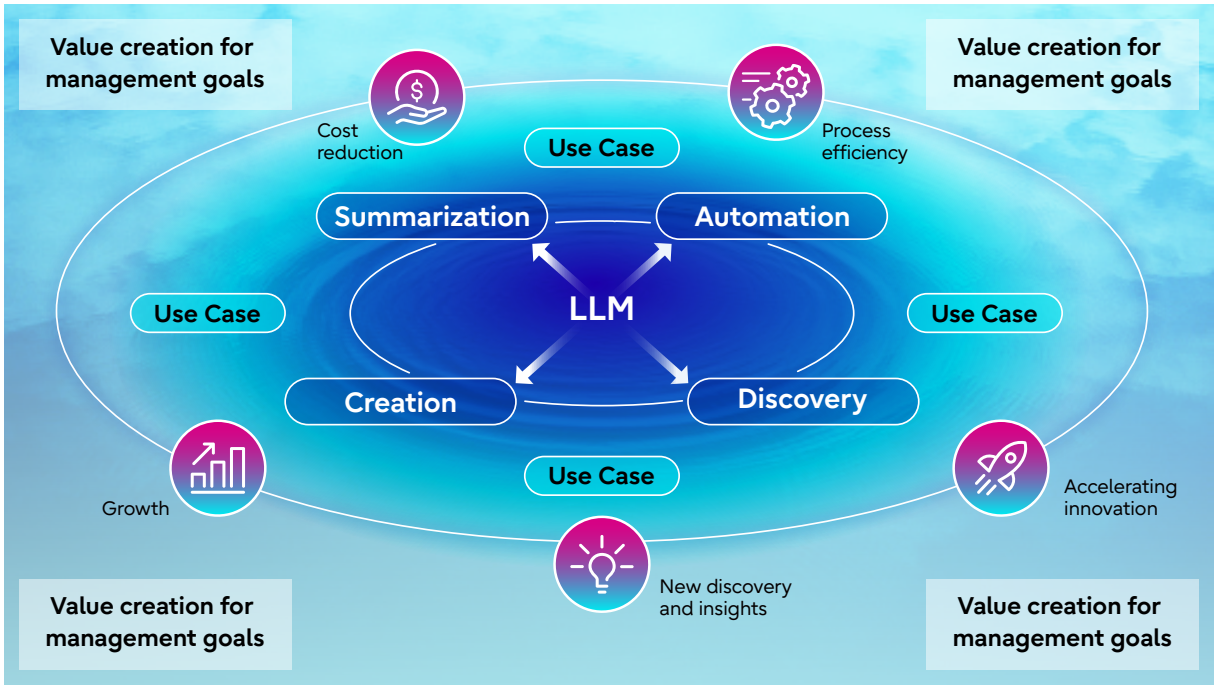
These are two important steps in the LLM lifecycle, model selection and optimization, organized by research and insights. However, it is important to understand that business owners are focused on maximizing value by looking at the entire LLM lifecycle. And it's up to management to realize the full value potential of LLMs.

(1) Recognize a new infrastructure that goes beyond machines and evolves "humanly" and leverages LLMs

What has become clear from the previous insights is that LLMs, unlike Conventional AI, do not address each step of the business (individual tasks), but instead becomes an intelligent engine that can address multiple steps of the business at once, transcending machines and becoming a new "human" evolving infrastructure. This is an evolution from traditional physical infrastructure and software as a machine.

As shown in Figure 7, LLMs are a driver of value creation, participating in the value creation process through use cases that businesses can feel. This "human" infrastructure is impacting all industries, making them more accessible and creating a productivity revolution for people. The level of automation has also evolved from Conventional AI RPA to EPA to CA. In other words, where humans were previously directly involved in processes and decisions (in the loop), the AI that powers LLM will take control, and humans will act as observers and final approvers (on the loop). In addition, LLMs have the potential to augment human innovation capabilities and navigate the future of business.

Figure 7 From LLM to Value Creation: The Evolution of Generative AI



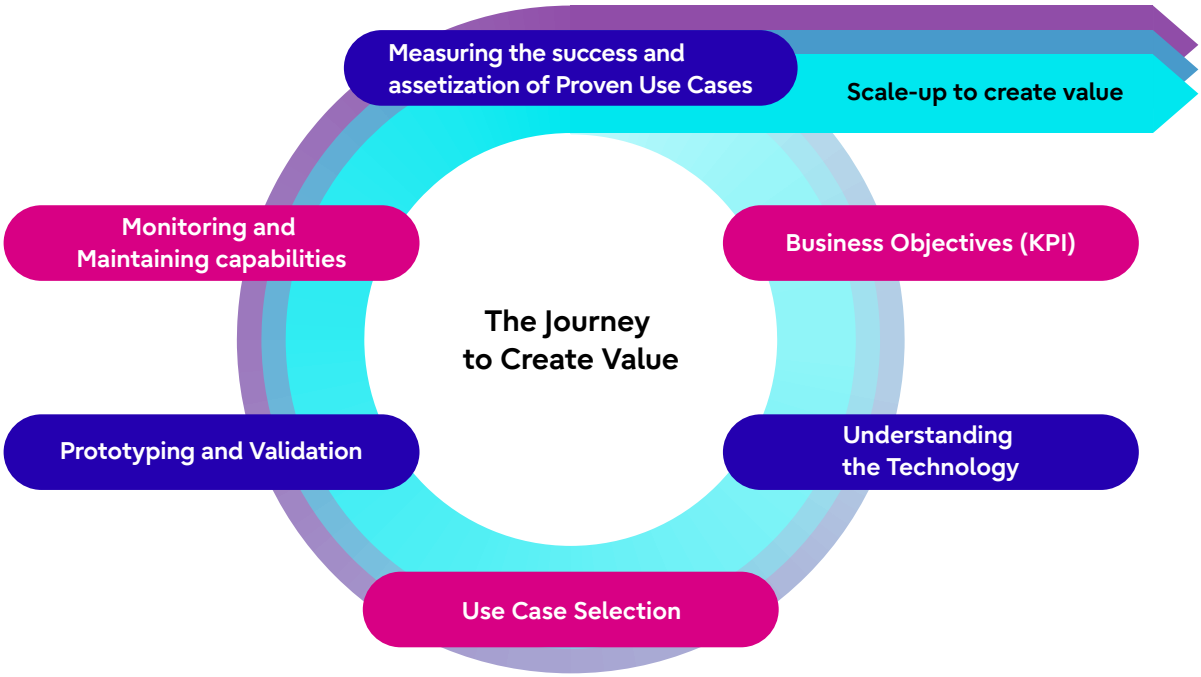
Source: Created by the author

(2) Use LLMs to maximize management performance: From Top Line to Bottom Line

The adoption of generative AI (LLM) in organizations has been identified as a top new technology investment strategy in many management surveys. However, many organizations are still stuck in the proof-of-concept (PoC) and use case establishment phases, and their management impact (contribution to management performance) is not significant. This has been described as "pilot purgatory"[*25] and "use case death".[*26]

As shown in Figure 8, the implementation of generative AI (LLM) requires validation and validation of PoC and use cases in the new value creation process. However, technologies, know-how, and use cases (such as architecture) that have been proven through repeated practice require rapid scale-up through capitalization. During the scale-up phase, it is possible to skip the PoC and use case demonstrations and move directly to implementation. To maximize management performance, it is important to take an approach that simultaneously achieves the top line (growth target) and the bottom line (profit target). Otherwise, you may lose support in your organization and your company's growth may stagnate.

Figure 8 Unleashing the power of generative AI: effective steps and scaling



Source: Created by the author

*25 WEF (December 2023) "Global Lighthouse Network: Adopting AI at Speed and Scale"
*26 McKinsey Digital (July 2023) "Technology's generational moment with generative AI: A CIO and CTO guide"

(3) Addressing potential mixed risks associated with LLM operations

As companies race to develop and deploy generative AI solutions, more and more of the risks that come with LLMs are being identified and mitigated through a variety of approaches. The hallucinations, toxicity (discriminatory output, etc.), potential intellectual property infringement, and poisoning (data and prompt attacks) associated with LLMs have been recognized as risks in response, research and countermeasures are being actively pursued, primarily by LLM vendors, and have been successful in moderating these risks. For example, RAG has proven effective in solving model hallucination problems by providing fact-checking and access to up-to-date information.

However, as the use of LLMs continues to grow rapidly, an organization's over-reliance on LLMs can create a new risk of disruption to the business itself in the event of a system accident or LLM failure. In fact, it has been reported that major generative AI companies have experienced system failures.*27 In contrast, some advanced companies have taken measures such as implementing multiple models.*28 In addition, using commercial LLMs as a core part of external solutions can undermine your competitive advantage.

Finally, external risks can be captured within LLMs when using external services via RAG or API. Internalizing external risk can be a serious problem, especially as we expect to see more external services via APIs in the future.

*27 OpenAI Shut Down ChatGPT to Fix Bug Exposing User Chat Titles; ChatGPT suffers the biggest system crash.
*28 Jasper (2023) "Jasper vs. ChatGPT: Which is Right For You?"

About the author

Dr. Jianmin Jin

2020 Fujitsu Ltd., Chief Digital Economist
1998 Fujitsu Research Institute, Senior Fellow
Dr. Jin's research mainly focuses on global economic, digital innovation/ digital transformation, and Dr. Jin has published books such as "Free Trade and Environmental Protection", etc.
Recent writings: the following Fujitsu Insight Paper, etc.

- Generative AI: Use Cases as the Pathway to Value Creation (2024)
- Transformative Quantum Computing: Striving for Greater Heights in Pursuit of Steady Progress (2023, Co-author)
- Transforming Supply Chains to Be More Productive, Resilient, and Sustainable (2023)
- Transformative Enterprise 5G: To Become an Attractive Enabler for DX (2023)
- The Composable Enterprise Emerging in the VUCA Era: From Concept to Practice (2023)

The author wishes to express profound gratitude to Shinichi Komeda, Hiroshi Nishikawa, Yasutoshi Kotaka, Naomi Hadatsuki, and Martin Schulz for their insightful discussions and invaluable advice during the development of this paper. Further appreciation is extended to Yoshihiro Mizuno for his guidance, Hiroshi Nishikawa, Nick Cowell, Bryan McMahon, Naomi Hadatsuki, Shinichi Komeda, and Takashi Shinden for their diligent review and native checks amidst their busy schedules. Lastly, heartfelt thanks to Michiyo Hano, Yukiko Sato, and Mitsuo Tsukahara for their unwavering daily support.

FUJITSU

February, 2024 v1.0