appliedAI
institute
for europe

This image was artificially generated by the appliedAI Institute for Europe gGmbH using the AI System Midjourney.

AI INSIGHTS

Making Business Decisions in the Realm of Large Language Models

Eine gemeinsame
Initiative

UNTER
NEHMER
TUM

Ipai

Content

1. Deriving Value Creation from LLMs 2

2. Understanding the LLM Ecosystem to prepare Business Decisions 4

2.1. Required Components for a flourishing LLM Ecosystem 4

2.2. When the iPhone vs. Android Moment comes - The Pros and Cons of open-source vs. closed-source LLM Alternatives 7

3. Supporting critical Business Decisions in Leveraging Large Language Models 10

3.1. The LLM Tech Stack 10

3.2. Key Influence Factors for LLM Make-or-Buy Decisions 12

3.3. Six Approaches for LLM Make-or-Buy Decisions 17

3.4. Perspective: Vertical Foundation Models 21

4. Conclusion 22

The appliedAI Institute for Europe gGmbH is supported by the KI-Stiftung Heilbronn gGmbH. The appliedAI Institute for Europe gGmbH is a subsidiary of the appliedAI Initiative GmbH.

1. Deriving Value Creation from LLMs

At the beginning of this decade, only a handful of tech enthusiasts and visionaries had ever heard of "generative AI". Today, only very few years - respectively even months - later, it seems clear that generative AI, particularly techniques related to Large Language Models (LLMs), already constitutes a major game-changer for individuals, businesses, and society as a whole. Indeed, without being overly dramatic, the significance of generative AI and LLMs can hardly be overestimated.

By enabling the automation of many tasks that were previously performed by humans, generative AI will significantly increase efficiency and productivity across entire value chains and corporate functions, thereby reducing costs as well as opening up new and exciting opportunities for growth. Based on the advances in Natural Language Processing (NLP), the underlying user interaction and experience with applications have changed as non-experts can now use these applications through providing instructions via natural language instead of technical code. Indeed, with generative AI, computers can now arguably innovate. They can produce novel content in response to prompts, draw from data they've ingested and interact with users.

The potential of generative AI within business is hereby twofold. Firstly, generative AI has the potential to transform entire business processes as LLMs possess capabilities to retrieve knowledge based on tasks and context, to understand the task and identify its solution, and eventually to make decisions for executing solutions along processes. Secondly, generative AI can incrementally advance existing products as well as develop new products such as automatically write blogs, sketch package designs, write computer code, compose and synthesize music, come up with new product ideas, or even theorize on the reason for a production error. And these are only a few of the potential use cases generative AI is suitable for.

The strength of foundation models, and therefore modern generative AI, lies in their ability to be tailored for various applications. In certain cases, minimal fine-tuning is needed with limited data to perform the desired tasks. However, in other cases, tasks can be accomplished by simply providing instructions without any examples (known as zero-shot learning) or with a small number of examples (known as few-shot learning). By utilizing and customizing foundation models, developers can now create AI applications that would have been impossible without recent advancements. These applications include interactive assistance tools (commonly referred to as co-pilots) for software development, as well as content generation and enhancement across different domains and modalities.

These developments allow (almost) everybody to use and benefit from AI, leading to enormous opportunities for value creation and productivity boosts. Short term, especially the enterprise functions (1) marketing / sales, (2) customer operations, (3) research and development, and (4) software engineering will benefit most from implementing generative AI. A study by McKinsey estimates that generative AI could add value between $2.6 trillion and $4.4 trillion annually and automate work activities that currently account for approximately 60-70% of employees' time[1]. This, in turn, leads to the conclusion that firms not employing AI are likely to be left behind.

[1] McKinsey and Company (2023). The economic potential of generative AI: The next productivity frontier. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/The-economic-potential-of-generative-AI-The-next-productivity-frontier#business-and-society.

Specifically referring to LLMs, research shows that approximately 15% of all tasks - i.e. also tasks traditionally performed in white-collar jobs - could be completed significantly faster when employing LLMS while maintaining the level of quality of non-automatized task completion. This proportion even rises to approximately 47 - 56% of all tasks when incorporating software and tooling built on top of LLMs employed. Moreover, the utilization of LLMs enables firms to transform their internal knowledge management systems. As such, employees can easily retrieve stored information by posing questions and engaging in dialogue, which significantly accelerates the overall decision-making process.

Overall, this AI insight article pursues a rather non-technical perspective with the aim to provide professionals in profit and non-profit organizations with the necessary insights and impetus to effectively navigate the realm of LLMs when making business decisions. By understanding the LLM ecosystem, professionals can derive value creation from LLMs and support critical business decisions. Additionally, this AI insight article explores the LLM tech stack and presents six approaches for the make-or-buy decision regarding LLMs. By following these guidelines, professionals can harness the power of this revolutionary technology and learn about how to integrate LLMs into their firm.

Glossary

Generative AI: A field of artificial intelligence that focuses on creating models capable of generating new content, such as text, code, images, or music, that resembles human-created content. Generative AI also aims to enable machines to exhibit creativity and autonomy in producing diverse and original outputs, often leveraging techniques like neural networks and probabilistic models to foster innovation and adaptability

Foundation Model: A large neural network model that captures and generalizes knowledge from massive data and serves as a strong starting point for further customization as well as a fundamental building block for various specific downstream tasks, empowering developers and researchers to efficiently develop a wide range of applications across different domains.

Large Language Model: A powerful neural network algorithm designed to understand and generate human-like language, typically trained on a vast amount of text data and considered a type of foundation models.

Pre-training: The initial phase of training a neural network model where it learns from a large dataset, allowing the model to capture general knowledge and patterns from the data to enhance its performance and adaptability.

Fine-tuning: The process of adapting a pre-trained neural network model to perform specific tasks by training it on task-specific data, allowing the model to specialize its knowledge and improving its performance on specific applications.

2.

Understanding the LLM Ecosystem to prepare Business Decisions

2.1. Required Components for a flourishing LLM Ecosystem

In order to enable firms to employ LLMs it is essential to be embedded in a larger LLM ecosystem. Establishing such a comprehensive LLM ecosystem requires several key components:

Computational infrastructure

The first key component for a LLM ecosystem in Europe is the availability of substantial computational infrastructure. LLMs require significant computational power to train and deploy effectively. Training a language model involves processing vast amounts of data and performing complex mathematical operations. Additionally, storage systems are required to store the large volumes of training data and trained models. This storage infrastructure needs to be scalable, reliable, and capable of handling processing of large amounts of data. Furthermore, distributed computing frameworks (such as Apache Hadoop or Apache Spark) are required for parallel processing and distributed training of LLMs. These frameworks enable efficient utilization of computational resources across multiple machines, reducing training time and improving scalability. In addition to the hardware infrastructure, the availability of specialized software and libraries is critical. This includes frameworks such as TensorFlow, PyTorch, or Hugging Face's Transformers, which provide the necessary tools for building, training, and deploying LLMs. Collaborations between research institutions, industry, governmental bodies, and cloud service providers can foster the availability of the required computational infrastructure, enabling Europe to establish a strong LLM ecosystem and drive advancements in natural language processing.

Pre-trained models

Generally, pre-trained models are essential components of LLMs as they serve as the initial building blocks for their development and functionality. These models are pre-trained on large datasets, enabling them to capture a general understanding of language, grammar, and semantic relationships. By starting with a pre-trained model, LLMs can leverage this knowledge and transfer it to specific tasks and reduce the need for extensive training on domain-specific data. Pre-trained models provide a high baseline performance due to their training on diverse and extensive datasets. They possess a broad understanding of language, allowing them to generate coherent and contextually relevant responses. Fine-tuning pre-trained models further enhances their performance, as they learn to adapt and specialize for specific tasks or domains. Additionally, pre-trained models offer versatility and generalization capabilities. Their broad knowledge of language allows LLMs to handle a wide variety of inputs and perform comparatively well across different industries, verticals and applications. This versatility enables firms to apply LLMs to various use cases, ranging from text generation and translation to sentiment analysis and summarization.

Training data

Training data is fundamental for LLMs as it forms the basis for learning patterns, understanding context, as well as generating accurate and contextually relevant outputs. Thus, to develop LLMs, access to a wide range of data sources is necessary. Collaborative efforts among industry, academia, and public institutions can facilitate the acquisition and sharing of training data while addressing privacy and legal considerations. Data sharing initiatives, such as research partnerships or data consortiums, can be established to pool resources and make large-scale, multilingual training datasets accessible. Furthermore, initiatives to collect and curate domain-specific datasets are necessary to enhance the deployment and accuracy of LLMs in specific application areas. For instance, collaborations with healthcare institutions can provide access to medical texts, enabling LLMs to better understand medical terminology and generate contextually appropriate responses in the healthcare domain.

Talent

To establish a LLM ecosystem, it is essential to address the need for highly skilled professionals with expertise in e.g., data science, machine learning, and domain-specific knowledge. Currently, there is a vast shortage of such professionals and proactive steps need to be taken to attract new talent, retain existing talent and technically educate society as a whole. To attract talent, creating a favorable environment for researchers and professionals with a climate for innovation and risk-seeking is pivotal. Investing in large R&D projects, providing funding opportunities, setting up unbureaucratic employment processes, and offering competitive salaries and benefits are essential components to attract top talent. In turn, to retain already existing talent, it is essential to create an attractive and supportive environment. For example, partnerships between academia and industry can offer internship programs, mentorship opportunities, and joint research projects, allowing professionals to gain valuable experience and contribute to real-world LLM deployments. Moreover, investments in education and training programs are crucial. As such, universities and research institutions should offer specialized courses and degree programs in data science, machine learning, and NLP. Collaboration between academia and industry can further enhance these programs by incorporating real-world challenges and industry best practices.

Regulation

To differentiate itself from competing countries and to establish a responsible development of LLMs, Europe needs to establish a supportive regulatory framework that addresses legal and ethical aspects while also enabling firms to maintain innovativeness. Per definition, LLMs rely on large datasets, which can include personal and sensitive information. As such, it is essential to have robust data protection regulations in place, such as the General Data Protection Regulation (GDPR) in Europe, to establish transparency for LLM developers and users. Key issues may comprise data minimization, explainability, and user consent, to build trust with users and protect their personal information. Moreover, fairness and bias mitigation are also critical considerations. LLMs can inadvertently perpetuate biases present in training data, leading to discriminatory or unfair outcomes. Europe should therefore develop a regulatory framework that enables the development of trustworthy LLM applications. For example, explainability techniques, such as interpretable model architectures and explainable AI methods, can provide insights into how LLMs arrive at their decisions, helping users understand and trust the outputs generated by these models.

Financing

Establishing a European answer in line with the idiosyncratic European requirements (for example regarding explainability, trustworthiness, values, or privacy) to the currently predominantly non-European actors requires significant financial resources. As such, public funding in terms of specialized incentivization programs to de-risk the significant investments required are necessary. Establishing such public funding programs would also send a strong signal within Europe and incentivize startups, SMEs and large corporations to further engage in the entrepreneurial endeavor of establishing a LLM ecosystem. In this context, it is important to note that the timelines of such funding schemes need to be adapted to the extremely high pace in the LLM environment, i.e. grants must be allowed and distributed quickly and unbureaucratic. Alternatively, private financing or public-private partnerships are possible options for establishing a European LLM ecosystem. However, also in these cases it is essential that such initiatives are supported by governmental institutions. Such support may include that for example public institutions employing LLMs are legally required to rely on LLMs that are in line with specific European requirements.

Competence centers

Competence centers for LLMs are indispensable for driving scientific research, strengthening industrial competitiveness, addressing the skills gap, promoting economic growth and innovation, and fostering trustworthy AI practices in line with European values and regulations. By investing in and supporting the establishment of these centers, governments, academia, and industry can manifest Europe's position as a leader in AI research and development while simultaneously ensuring Europe’s digital sovereignty. This not only brings enormous economic advantages but also facilitates the responsible and beneficial application of AI technologies. Currently, the major European competence centers are predominantly focused in Germany (such as Laion, LEAM, or the appliedAI Institute for Europe), France (such as Hub France IA) and Sweden (such as AI Sweden). Indeed, a very limited number of competence centers should be established which provide cutting-edge research and are at the world's forefront of developing foundational models. These lighthouse initiatives should not only engage in topnotch research but also in actively transferring knowledge to the industry and training domain-specific foundation models to serve specific industrial use cases. Moreover, these competence centers may act as think tanks actively pursuing political advisory to facilitate a well-informed political decision-making that is based on well-proven facts and is in line with European values and regulations, especially regarding trustworthiness and risk mitigation.

Overall, it is essential to address these high-level points for building up a LLM ecosystem in Europe. Importantly, from a European perspective and due to the strong dominance of non-European actors in the current LLM environment, it is crucial that a decision to build up a European LLM ecosystem needs to be made quickly in order to not fall behind even further other competing actors (i.e. mainly tech firms from the USA and China).

2.2. When the iPhone vs. Android Moment comes - The Pros and Cons of open-source vs. closed-source LLM Alternatives

Recognizing historic moments while being part of them is a truly difficult task. However, the ongoing fourth industrial revolution and ongoing dissemination of LLMs are more than likely to enter future history books. ChatGPT, developed by OpenAI, achieved the milestone of reaching one million users in record-breaking five days after its launch on November 30, 2022. In comparison, it took Facebook approximately ten months to reach one million users after its launch in February 2004. The rapid adoption of ChatGPT in specific and LLMs in general underscores the enormous demand and interest in AI-powered agents and highlights the transformative potential of LLMs for a seemingless infinite number of application areas.

Indeed, the rapid explosion in adoption of ChatGPT concomitant with its novel capabilities can be compared to the "iPhone moment". Of course, ChatGPT was not the first LLM available. However, ChatGPT was the first LLM that was used not only by very few tech enthusiasts but also commercialized in a sense that it received broad attention by the general public. Just as Apple's iPhone revolutionized the smartphone industry in 2007, LLMs and generative AI revolutionize NLP, communication, and content generation. As such, the "iPhone moment for LLMs" depicts a significant shift in how humanity can interact with AI-powered technologies. Thus, the "iPhone moment for LLMs" represents a pivotal point in the development and adoption of generative AI as it is suddenly becoming widely both recognized and utilized. At the same time, not only the possibilities respectively capabilities but also the potential limitations of LLMs are becoming visible.

At present, the prevailing market landscape is predominantly ruled by closed-source, API-based language model systems. Nevertheless, the array of choices for open-source LLMs is steadily expanding. Figure 1 below provides an evolutionary overview of LLMs by tracing the development of LLMs from 2018 to 2023 and highlights some of the key models:

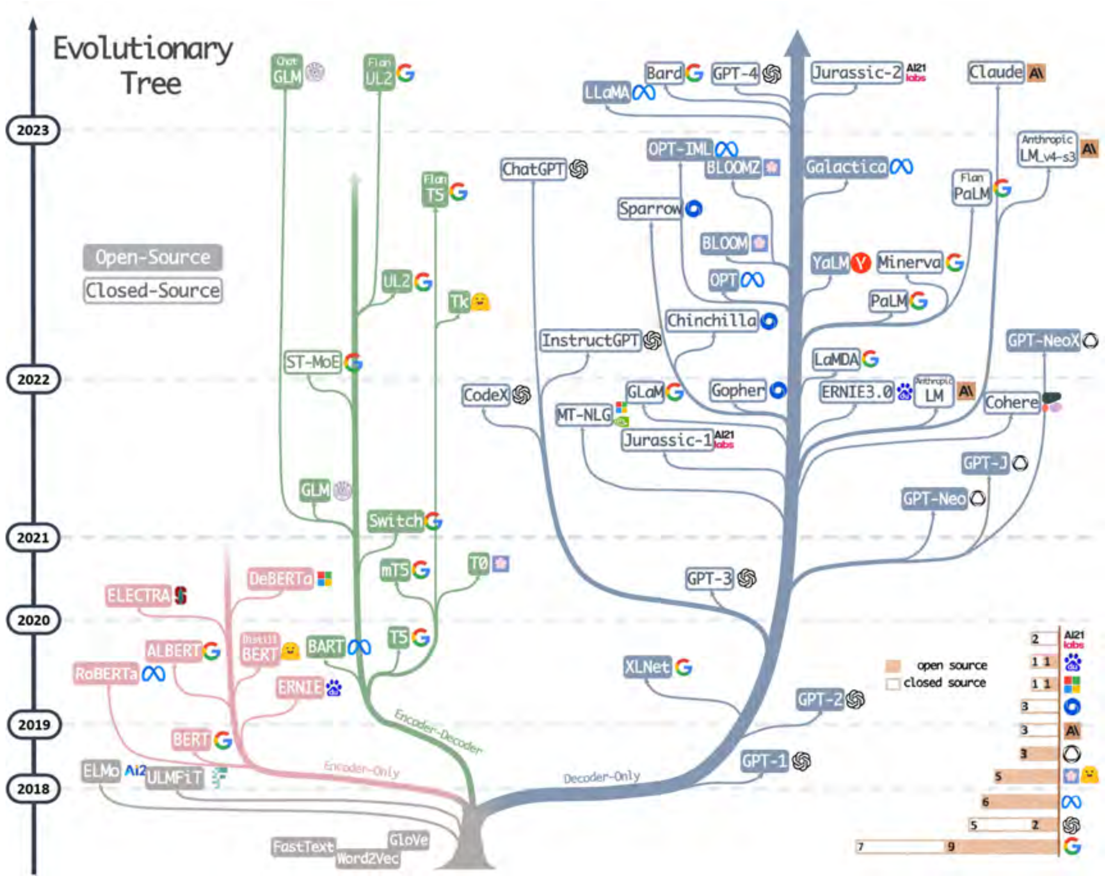

Figure 1: The evolutionary tree of modern LLMs. [2]

Generally, it needs to be carefully evaluated whether to employ open-source or closed-sourced models. Open-source LLMs provide several advantages over their proprietary counterparts. Firstly, they allow for greater transparency and auditability. Open-source models grant researchers and developers unrestricted access to the underlying code, model architecture, and training data, empowering them to gain insights into the model's inner workings and detect any potential biases or ethical considerations. Transparency stands as a pivotal aspect of open-source LLMs. By openly sharing the model's code and data, developers can thoroughly examine and validate the model's behavior, ensuring it adheres to the desired ethical standards. This level of transparency also addresses concerns about algorithmic biases and discriminatory outputs. Through collaborative efforts, researchers and the broader community can actively collaborate to identify and rectify these issues, fostering the development of more reliable and equitable language models.

Secondly, open-source LLMs foster collaboration and innovation. The open-source philosophy encourages a diverse range of researchers and developers to contribute their expertise to improve the models. This collaborative approach leads to the development of more robust and effective language models. By leveraging the collective intelligence and efforts of the community, open-source LLMs can evolve and address the challenges and limitations of the field more rapidly.

[2] Source:Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. arXiv preprint arXiv:2304.13712. https://arxiv.org/pdf/2304.13712.pdf.

Furthermore, open-source LLMs empower developers by providing them with the tools and resources to build upon existing models. The availability of pre-trained models and open-source libraries significantly reduces the barrier to entry for researchers and developers who want to experiment with language models or build applications on top of them. This accessibility fosters innovation and accelerates the pace of advancement in the field.

Indeed, several prominent open-source LLM initiatives have emerged, each making significant contributions to the field. Besides OpenAI's GPT (Generative Pre-trained Transformer), another influential open-source LLM initiative is Hugging Face's Transformers library. This library provides a comprehensive set of pre-trained models, including various architectures such as GPT, BERT, and RoBERTa. The library also offers tools and utilities for training, fine-tuning, and deploying models, making it easier for developers to leverage the power of LLMs in their applications. The Transformers library has gained widespread popularity due to its user-friendly interface, extensive documentation, and vibrant community support. In addition to OpenAI and Hugging Face, there are several other open-source LLM projects and libraries, such as Fairseq, Tensor2Tensor, and AllenNLP.

In contrast, closed-source LLMs offer a range of benefits that have contributed to their popularity among service providers and organizations. Firstly, these models often leverage significant computational resources and proprietary datasets during their training, allowing them to achieve high levels of performance on various language tasks. The investment in infrastructure and data acquisition by these companies can result in LLMs that surpass the capabilities of open-source models. Moreover, closed-source LLMs are typically developed and fine-tuned specifically for the company's own applications and services. This customization allows companies to optimize the models for their specific use cases, resulting in improved performance and tailored experiences for their users. Closed-source LLMs also provide companies with a competitive advantage, as they can differentiate their services based on the unique capabilities of their language models.

However, closed-source LLMs come with certain limitations and challenges. The lack of transparency and accessibility can hinder the ability of external researchers and developers to understand and evaluate the models. Without access to the underlying code, training data, or model architecture, it becomes challenging to identify potential biases, ethical concerns, or limitations within these models. Additionally, the closed nature of these LLMs restricts collaboration and innovation. External researchers and developers may not be able to contribute to the improvement or enhancement of the models, limiting the scope of advancement in the field. This closed ecosystem can lead to a concentration of power and restrict the availability of cutting-edge language models to a select few organizations or service providers.

3. Supporting critical Business Decisions in Leveraging Large Language Models

3.1. The LLM Tech Stack

In the context of integrating LLMs into business operations, a diverse range of choices must be carefully considered. These choices span from adopting external closed-source models through APIs to developing LLMs internally, with gradual approaches in between. Finding the most suitable approach for these make-or-buy decisions is not a straightforward task and requires a comprehensive evaluation of various factors. It is essential to consider both the LLMs themselves and their potential applications, extending the decision-making process beyond individual applications to encompass the broader implications of LLM utilization within the organization.

To attain this goal, the initial stage involves evaluating the available capabilities and internal resources to determine the appropriate tech stack to address. Typically, the LLM tech stack comprises the following layers:

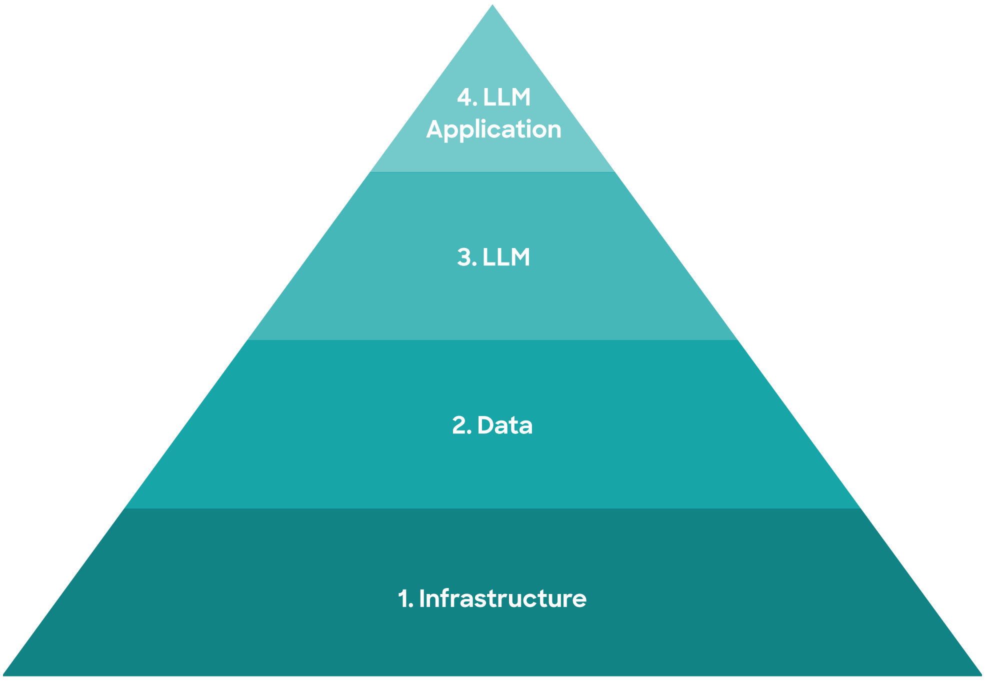

Figure 2: The tech stack for LLMs

The bottom layer refers to the **infrastructure** that is required (such as the necessary hardware or cloud platforms). This includes systems and processes for developing, training and running LLMs including high performance computation (HPC) optimized for AI computation, most specifically deep learning). The expected specific use cases and their scalability influence the overall infrastructure decision.

The second layer pertains to the essential **data** in terms of both quantity and quality. The volume of data needed is highly influenced by the method of utilizing and adapting LLMs, such as pre-training versus fine-tuning. Ensuring data quality and curation is of utmost importance for the effectiveness of LLMs. Businesses have the option to allocate resources towards data curation and preprocessing techniques, encompassing tasks like data cleaning, normalization, and augmentation, all of which serve to elevate data quality and coherence. By implementing stringent quality control measures during data collection and labeling, enhanced data reliability can also be achieved.

The third layer is centered around the core **LLM** itself, which can come in either open-source or closed-source forms and serves as the foundation for various unique applications. To establish a systematic make-or-buy strategy, companies should strive to create synergies among value-adding use cases.

Lastly, the fourth layer encompasses the actual LLM **applications** that make use of the language model. These applications can be developed as end-to-end solutions or rely on external third-party APIs. The decision to build or buy an application depends on the underlying layers. For instance, if a company lacks access to high-quality data, choosing the "make" option for application development may not be a feasible choice.

3.2. Key Influence Factors for LLM Make-or-Buy Decisions

Overall, there are eleven factors that need to be considered when being faced with a make-or-buy decision for LLMs:

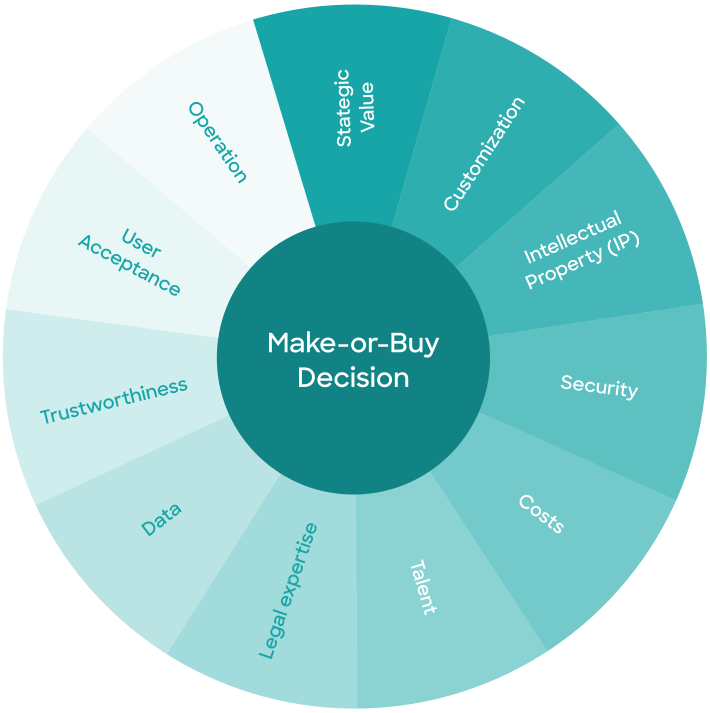

Figure 3: Key influence factors for LLM make-or-buy decisions

1. Strategic value

Clearly, ensuring that the deployment of LLMs is in line with the overall corporate strategy is of utmost importance for the make-or-buy decision dilemma. Put differently, the main reason for developing a LLM in-house is that it is of a high strategic value with a high scalability and value creation, i.e. enabling a firm to achieve (sustainable) competitive advantage. By building LLMs internally, organizations can establish and maintain proprietary knowledge and in-house expertise, creating a valuable intellectual asset. This intellectual property can contribute to long-term competitive advantage, as it becomes increasingly difficult for competitors to replicate or imitate. Regarding LLM fine-tuning, based on the quality and value of the training data competitive advantage may also be achieved. As fine-tuning approaches are comparatively inexpensive, this presents a promising opportunity to create value for firms with valuable data assets. In turn, when LLMs are bought (i.e. developed and trained) externally, they are available on the broad market and available to competitors as well - and thus, no sustainable competitive advantage can be reached. Moreover, having in-house LLM development capabilities fosters innovation and a culture of continuous learning, i.e. it enables firms to stay at the forefront of technological advancements.

2. Customization

Opting for in-house development of LLMs generally provides a higher level of customization, enabling tailoring the language model to specific requirements and use cases unique to the firm. This advantage holds true to some extent even when fine-tuning models using proprietary internal data. Customized LLMs offer greater flexibility and complete ownership compared to off-the-shelf products. On the other hand, non-customized external LLMs come with lower costs. However, it is important to note that potentially sensitive data must be shared with the external partner.

3. Intellectual Property (IP)

LLMs undergo training on extensive datasets that may contain copyrighted materials or proprietary information, particularly if sourced from the external market. Consequently, concerns arise regarding the ownership and usage rights of the content generated by these models. Establishing clear policies and agreements becomes paramount for companies to address IP rights concerning LLM-generated content. These guidelines should define content ownership, any licensing or usage restrictions, and provisions for safeguarding sensitive information. When collaborating with third parties, these considerations must be incorporated into the contracting process. It is crucial to acknowledge, however, that there remains significant uncertainty regarding IP rights concerning content produced through generative AI.

4. Security

LLMs undergo training on vast amounts of textual data, which might encompass confidential, proprietary, or personally identifiable information. Depending on the particular use case, LLMs may involve the processing of highly sensitive business data. Therefore, for each use case, it is essential for firms to conduct a comprehensive risk assessment in advance to identify and address potential security issues. For particularly sensitive data, hosting the LLM within the firm's isolated network is often advantageous. When this is not feasible, collaborating with reputable external LLM providers who adhere to strict security standards and offer transparency about their security practices becomes crucial. Additionally, for data falling under the purview of GDPR, firms must ensure that all data is stored and processed on servers located within Europe. This is to comply with the data protection regulations and safeguard the privacy of individuals.

5. Costs

Developing LLMs in-house is a very costly endeavor. Firstly, it requires significant investments in terms of hiring and paying a highly-skilled workforce, such as ML engineers and NLP specialists, who typically need to be attracted with high salaries. Secondly, the development process itself is both time-consuming and resource-intensive. It involves extensive research, data collection, model training, and iterative improvement cycles, which demand considerable computing power and infrastructure investments. Moreover, ongoing maintenance, updates, licenses, and support require continuous investments to ensure optimal performance and reliability. Lastly, it is important to consider the opportunity cost of allocating internal resources to LLM development instead of focusing on the core business activities. While in-house development offers several benefits, it diverts attention and resources from other strategic initiatives and potentially delays time-to-market, which in turn may lead to increasing opportunity costs. Therefore, executives should carefully evaluate the financial implications and weigh the costs against the potential benefits before deciding to develop LLMs in-house. In many cases, fine-tuning may be a more suitable approach with substantially lower costs. Moreover, to address the high development costs, organizations can explore ways to streamline the labeling and development cycles. Leveraging pre-existing labeled datasets or partnering with external data providers can reduce the need for extensive manual labeling, thereby saving time and resources. Additionally, adopting cloud-based solutions for data storage and processing can offer scalability and cost-efficiency, enabling organizations to handle large volumes of data effectively.

6. Talent

The scarcity and high demand for experienced professionals in fields such as data science, ML, or NLP often make it difficult to establish a skilled in-house team, especially for SMEs who are often confronted with resource constraints. Especially in Europe, the competition for top talent is fierce, and both SMEs and large firms are facing recruitment difficulties and talent shortage. Additionally, the extremely rapid development in the field of LLMs requires continuous learning and professional development, making it essential for companies to make significant investments in training and upskilling their existing workforce. Overcoming these talent-related hurdles demands a strategic approach, including fostering partnerships with academic institutions, collaborating with external partners, competitive salaries, and creating a stimulating work environment that promotes innovation. Thus, firms that are already confronted with a scarcity of talent may decide to source their LLM solutions from the market to save both direct and indirect talent-related costs and to utilize their restricted talent resources for other projects, respectively for strategically important firm-specific solutions that may not be bought from the market. In-house fine-tuning models often constitutes a middle course for striking the balance between acquiring off-the-shelf products and developing models from scratch.

7. Legal expertise

When developing LLMs in-house, firms require legal expertise to navigate the increasingly complex regulatory landscape. For instance, the EU AI Act focuses on the prevention of harm to health, safety and fundamental human rights. By taking a risk-based approach, AI systems are assigned to a risk class where high-risk systems - such as LLMs - must meet stricter requirements than low-risk requirements. Thus, firms pursuing an in-house development of LLMs must ensure that they follow all regulatory requirements which requires an increasingly complex legal expertise. In case firms do not have this in-house legal expertise or want to reduce their general liability, they may rather decide to buy a LLM from the market and ensure that the provider is fully liable, i.e. that the specific use case is in line with applicable laws and regulations. Additionally, by considering the risk classification early in the decision-making process and making timely go / no-go decisions, firms can avoid unnecessary expenditures and undesired legal consequences.

8. Data

Data is of utmost importance for the performance of LLMs. LLMs rely on vast amounts of diverse data to understand language patterns, enhance accuracy, and generate coherent responses. This enables them to generate fluent and contextually appropriate responses, making interactions more natural and effective. However, biases present in the data can pose challenges. LLMs might inadvertently learn and perpetuate biases present in the training data. To address this, efforts are being made to identify and mitigate biases. Diverse and inclusive training data is crucial to ensure fairness and reduce the amplification of existing biases. For example, regular monitoring and user feedback are vital in detecting and rectifying biases. By evaluating LLM outputs and actively seeking user input, developers can improve the system's fairness and mitigate biases. Clearly, data is equally important for the process of fine-tuning LLMs. By fine-tuning on domain-specific data, LLMs can acquire specialized knowledge and language patterns related to the target task. This enables them to generate responses that align with the specific requirements of the use case. Moreover, fine-tuning also helps address biases and improve fairness in LLM responses. By fine-tuning with datasets that are explicitly designed to be diverse, inclusive, and representative, developers can reduce biases and ensure that the LLM performs more equitably.

9. Trustworthiness

Trustworthiness is of paramount importance when employing LLMs. The in-house development of LLMs allows firms to have full control over the entire process, enabling them to build the LLMs in line with their values and ethical considerations. This control fosters trust by ensuring that the LLMs are aligned with the firm's mission and vision. Moreover, in-house development provides transparency and explainability. As such, firms can document and communicate the development methodologies, data sources, and training processes, allowing users to understand and evaluate the LLMs' outputs. Moreover, by mitigating biases and ensuring fairness, firms can build trust among users, assuring them that the LLMs provide accurate and unbiased information. However, when buying LLMs from the market and especially from established suppliers, firms may benefit from the fact that the acquired LLM has undergone rigorous testing, evaluation, and compliance checks to ensure their LLMs meet industry standards and regulatory requirements. Again, the fine-tuning of models often constitutes a compromise between trustworthiness and effort.

10. User Acceptance

User acceptance issues are another concern when it comes to LLMs. User acceptance can be influenced by the impact on job roles and responsibilities. Employees may fear that LLMs could automate tasks or replace their roles. However, involving users early in the deployment process, providing training for upskilling, and emphasizing the collaborative potential of LLMs can help users see them as tools that enhance their productivity rather than threats to their job security. Another user acceptance issue is unrealistic customer expectations regarding the robustness and accuracy of the models can lead to dissatisfaction. Thus, from a management perspective it is crucial to set clear and well-defined acceptance criteria, ensuring that customers have a realistic understanding of what the models can and cannot accomplish. Effective communication and managing expectations play a vital role in ensuring user acceptance and satisfaction.

11. Operation

Operational considerations also come into play when implementing LLMs. One of the risks is a lack of LLM-related expertise across different operational stages, from data preprocessing and model training to deployment and maintenance. Organizations need to foster a multidisciplinary approach and ensure collaboration between NLP practitioners, data scientists, and domain experts to maximize operational efficiency and effectiveness. Furthermore, model performance issues such as concept drift or data drift require continuous monitoring and adaptation to maintain optimal results.

Clearly, all these factors cannot be viewed in isolation but rather must be acted upon taking a holistic approach. The development and deployment of LLMs are accompanied by several pain points and risks. The high development costs, data quality challenges, compliance with regulatory frameworks, security challenges, and operational considerations all demand careful attention and strategic planning. Overcoming these obstacles requires organizations to invest in resources, expertise, and infrastructure while ensuring adherence to ethical and legal guidelines. By addressing these pain points and mitigating the associated risks, firms can harness the power of LLMs to unlock valuable insights, enhance decision-making processes, and drive innovation in various industries. Strategies such as optimizing development cycles, improving data quality, adhering to regulatory requirements, and fostering a culture of continuous learning and adaptation can help organizations navigate these challenges and mitigate risks effectively.

3.3. Six Approaches for LLM Make-or-Buy Decisions

With the LLM tech stack and relevant key factors and business considerations in view, there are six generic approaches that firms can follow in LLM make-or-buy decisions:

1. Buy an end-to-end application without LLM controllability

When assessing use cases that offer low strategic value and have limited customization requirements for both the application and LLM, opting for a pre-built end-to-end application constitutes the preferred and convenient solution. In such cases, the LLM functions as a concealed component, seamlessly integrated into the application. Since the LLM is specifically tailored to the application's needs and scope, explicit customization and control over this hidden LLM are unnecessary and likely restricted by the vendor.

An example for this approach may refer to a firm that intends to leverage sentiment analysis to analyze customer feedback and thereby, to gain valuable insights for product improvement. By monitoring online reviews and social media comments the firm can gauge the sentiment of its customers towards their products and services via an end-to-end application. Sentiment analysis will provide the firm with a comprehensive understanding of the customers' positive and negative experiences. For instance, if a significant number of customers express dissatisfaction with a specific feature of a product through negative sentiments, the company can identify areas that need improvement. By using sentiment analysis, the firm can make data-driven decisions to enhance product quality and ultimately strengthen customer loyalty and satisfaction.

2. Buy an application with limitedly controllable LLM -Procure the application including LLM as a component with some transparency and control)

This approach of procuring an application along with controllable LLMs applies to use cases that demand minimal adjustments or can be deployed immediately. It is worth noting that in scenarios where customization needs are low, the significance of controlling the underlying LLM might be also minor, and companies may only focus on adapting the user layer to meet their specific requirements. Nevertheless, it is crucial to carefully consider use case-specific requirements concerning the degree of customization, regulation, data security/ secrecy, IP concerns, and overall performance. Another important consideration hereby is the reusability of a LLM across applications in the company that can yield to undesired dependencies and vendor-locking scenarios. In the end, this approach is only feasible in case of low data confidently allowing transfer to external providers.

For instance, the firm of the previous example now seeks to conduct a more targeted sentiment analysis specifically focussing on the firm’s offerings. The firm therefore decides to buy a pre-built application that comes with a limitedly controllable LLM (e.g., customization of prompts and prompt templates). Since the firm's primary goal is to make a more specific sentiment analysis taking into account the firm's idiosyncratic products, they opt for a pre-built solution.

3. Make application, buy controllable LLM – Internal development of application on top of procured LLMs via APIs

As an alternative to the approach mentioned above, a further option involves exclusively focusing on the internal development of the application while sourcing and integrating externally sourced pre-trained or fine-tuned LLMs. This approach is particularly suitable for use cases that demand medium to high levels of customization of an LLM, especially when internal resources such as computing power, capacity, or specific skills are not sufficiently available. Additionally, budget constraints can also drive the decision to adopt this strategy. However, the same consideration as in the second approach with regards to e.g., customization, regulation, data security / secrecy, or IP as well as the overall performance and model reusability need to be carefully taken into account, and vendors should be carefully analyzed.

For example, a firm specializing in natural language processing applications decides to develop a language translation service with advanced capabilities. Due to resource limitations in terms of computing power and costs, the external sourcing of pre-trained or fine-tuned LLMs constitutes a practical solution instead of building the entire system from scratch. By procuring the LLM via APIs, the firm can focus on the internal development of their language translation application without the need to train language models from the ground up.

4. Make application, fine-tune LLM – Internal development of application and fine-tuning of LLM based on procured or open-source pre-trained LLMs

This approach involves utilizing existing pre-trained LLM models, along with specific fine-tuning frameworks or services, and combining them with internal development efforts to build applications and fine-tuned models using internal data for targeted use cases. The quantity and quality of open-source pre-trained LLMs are continuously expanding, thanks to various providers and open-source initiatives. However, it is crucial to note that the licenses of these pre-trained models can impose significant limitations on their commercial use. When it comes to fine-tuning services, several providers such as AWS, Google, NVIDIA, H2O and others already offer such services. Additionally, various open-source fine-tuning services are already available as well. The level of internal development required depends on both the sophistication of the fine-tuning components and the quality of the underlying pre-trained LLM, as well as the availability of in-house data. While fine-tuning models is comparatively inexpensive, data quality often poses a major bottleneck. Nevertheless, this approach offers a viable option for achieving sufficient customization and quality of LLMs, all while maintaining control over internal data processing and hosting of the LLMs. This becomes particularly crucial in certain use cases, ensuring sustainable competitive advantage.

For instance, a medical research institution may aim to develop a specialized medical diagnosis system that can accurately identify rare diseases from patient symptoms. To achieve this, they decide to adopt the approach of making their application while fine-tuning a LLM. Specifically, the firm decides to source a pre-trained LLM from an open-source initiative that provides a solid foundation for NLP tasks. However, they recognize the need for specific fine-tuning to make the LLM more proficient in medical diagnosis. Using their extensive medical database, the firm then fine-tunes the pre-trained LLM by incorporating medical literature, patient records, and further in-house data to adapt it to the intricacies of rare disease identification. They leverage both internal development efforts and fine-tuning frameworks to achieve the desired level of customization. The availability of open-source fine-tuning services further aids their efforts, enabling them to efficiently optimize the LLM for their targeted use cases. By fine-tuning the LLM in-house, the firm maintains full control over their proprietary medical data, ensuring privacy and compliance with healthcare and data privacy regulations.

5. Make application, pre-train LLM - Internal development of application and pre-training of LLM from scratch

This approach focuses on a full end-to-end development ("make") and involves building the application itself as well as pre-training LLMs in-house from scratch. The broader the applicability of a LLM across various applications and the greater value it can generate, the more advisable it becomes to pursue the "make" approach. This also accounts for highly sensitive use cases where relying on externally sourced models is not an option. Although very costly, developing LLMs from scratch might be the best option for achieving optimal customization and quality of LLMs, and, therefore, for ensuring sustainable competitive advantage.

For instance, a medical technology firm may aim to create a state-of-the-art medical diagnosis assistant capable of accurately diagnosing various medical conditions based on patient symptoms, pathological results and medical history. To achieve this, they decide to adopt the approach of making their application while pre-training a LLM in-house from scratch. Specifically, due to the sensitive nature of medical data and the importance of data privacy, MedTech Solutions chooses not to rely on externally sourced models. Recognizing the critical nature of medical diagnoses and the need for a highly specialized LLM, the firm believes that developing a custom LLM will offer the highest level of accuracy and applicability across diverse medical conditions. By creating the LLM from scratch, they ensure complete control over the model's architecture, parameters, and training process, tailoring it precisely to the complexities of medical diagnosis. While the development of an LLM from scratch is a significant investment of time and resources, the firm views it as an essential step toward achieving optimal customization and quality for their medical diagnosis assistant. By pursuing this approach, they aim to secure a sustainable competitive advantage by providing accurate and reliable medical diagnoses, ultimately improving patient outcomes and healthcare practices.

6. Stop

If the use case holds limited strategic value, it is advisable to assign (limited) resources to use cases of higher strategic significance.

A general view on fine-tuning vs. pre-training models from scratch

Fine-tuning pre-trained LLMs generally incurs significantly lower costs compared to building LLMs completely from scratch. Indeed, depending on the underlying data structure and volume, fine-tuning costs are comparatively low, reaching from a few hundred dollars to a few thousand dollars. When fine-tuning, the pre-trained model is already available, eliminating the need for resource-intensive pre-training on vast amounts of data and computational power. This translates to significant savings in terms of computational resources, time, and electricity consumption.

Pre-training LLMs from scratch, in turn, involves substantial costs at various stages that can reach up to two-digit million dollar spheres. For example, the training costs for OpenAI's GPT-3 are estimated at approximately $5 million[3], while models with more training parameters are estimated to further exceed these costs. Indeed, pre-training LLMs from scratch demands an enormous amount of computational power, including specialized hardware and extensive infrastructure, making it an expensive endeavor. Additionally, the pre-training process can take weeks or even months to complete, further adding to the costs associated with computational resources and electricity.

Data acquisition and annotation costs are also considerably different. Fine-tuning LLMs typically requires a smaller labeled dataset for the target task, which can be less expensive to obtain, annotate, or curate compared to the comprehensive and diverse dataset required for pre-training an LLM from scratch. As such, the costs of acquiring and labeling a large-scale dataset can be substantial and requires substantial domain expertise paired with significant human effort.

Overall, fine-tuning LLMs generally offer cost advantages compared to pre-training LLMs from scratch. However, it is essential to consider the specific requirements of each specific use case, including the scale of the target task, availability of data, and potential risks, to determine the most appropriate approach based on the available resources and objectives. Clearly, the final decision on whether to pre-train LLMs from scratch versus fine-tuning available models also depends on the business cases and financial resources a firm is willing to invest.

[3] https://www.unite.ai/can-you-build-large-language-models-like-chatgpt-at-half-cost/

3.4. Perspective: Vertical Foundation Models

The current advancement of LLM development is characterized through an extremely high pace. However, one development that is still in a fledgling stage is the integration of vertical foundation models. Specifically, the integration of vertical foundation models refers to employing specified LLMs for specific industries. In this approach, foundation models - such as OpenAI's GPT-4, Aleph Alpha's Luminous or Meta's Llama 2 - are adapted on domain-specific data for domain and industry-specific applications in a first step (cf. Figure 4). This leads to having highly specialized vertical models combining open-domain with domain-specific knowledge. Subsequently, in a second step, further fine-tuning on application-specific data (such as text or image data) results in an application-based vertical model. The advantage of such a vertical model value chain lies in the very high degree of specialization compared to generally pre-trained models without a specific industry focus. Additionally, vertical models are crucial to scale LLM development as these models can be reused across applications of the respective application domain.

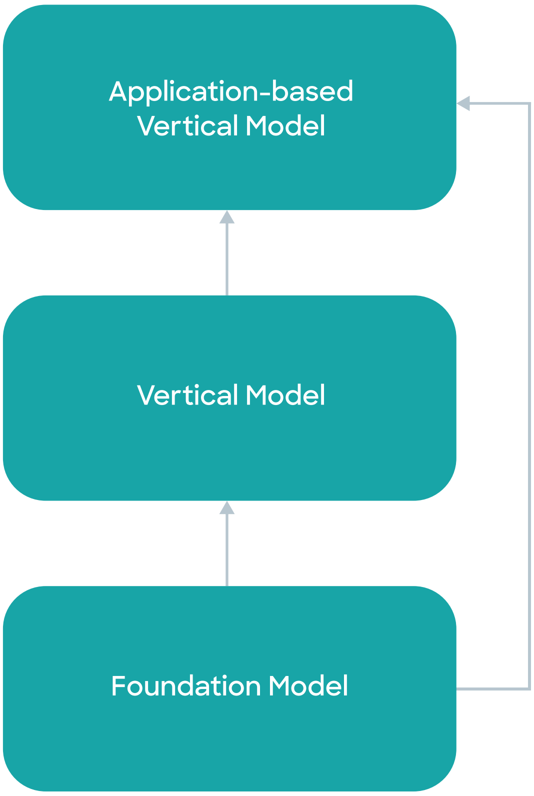

Figure 4: Vertical model value chain

4. Conclusion

LLMs will transform the way we work, significantly impact our jobs, change how we innovate, and alter how we generate content. Indeed, LLMs already have revolutionized language generation and various other tasks, showcasing their ability to generate coherent and contextually relevant output. These models leverage deep learning techniques and large-scale training data to learn patterns and semantic representations, enabling them to understand and generate human-like language. As such, it comes to no surprise that LLMs are predicted to significantly impact a whole range of tasks performed in a working environment in a way that they will be completed significantly faster while simultaneously maintaining the level of quality.

It is important to acknowledge that while LLMs have shown promising progress, there remain significant challenges that need to be addressed in the days to come. As mentioned earlier, one major challenge pertains to AI computational infrastructure, as leveraging LLMs in practical applications requires substantial computational resources and specialized infrastructure, an area where Europe currently faces limitations. Another challenge involves the curation of high-quality data, particularly in industries like manufacturing, utilities, and construction, where ensuring factual correctness and consistency is essential but often poses a bottleneck. Also, meeting European regulations - such as the EU AI Act - and avoiding biases also requires careful attention to data quality and representativeness. Furthermore, there are still challenges related to LLM algorithms per se, including, for example, the development of multilingual models that perform well across all languages and especially under-resourced ones, responsible reasoning in social contexts involving moral decision-making, and grounding LLMs with less studied modalities such as physiological, sensorial, or behavioral data. These challenges encompass various aspects of the pipeline, from data collection and representation to model design, evaluation, and deployment, and they highlight the need for continued research and innovation in the field of LLMs. By cooperating to address these challenges, we may fully leverage the capabilities of LLMs and ensure their responsible and impactful use in diverse applications.

In all, it is important for researchers, firms, and policymakers to closely collaborate to harness the potential of LLMs while addressing the associated challenges. By advancing LLM research and application in a responsible and ethical manner, we can unleash their full potential and leverage them as powerful tools that will change the way we live. Thus, now is the time to act and to actively prepare for the exciting opportunities and challenges that LLMs will bring to us.

Authors

Dr. Philip Hutchinson
p.hutchinson@appliedai-institute.de

Philip Hutchinson is an AI Analyst at appliedAI Institute for Europe gGmbH. He has long-standing experience in the field of AI and worked for Ernst & Young prior to joining appliedAI. Philip holds a PhD from Kiel University where he conducted research at the intersection of innovation management and Artificial Intelligence.

Bernhard Pflugfelder
b.pflugfelder@appliedai.de

Bernhard Pflugfelder works as Head of Use Cases and Applications at the appliedAI Initiative GmbH. Bernhard has 15 years of experience in the fields of Data Science, Natural Language Processing (NLP), as well as data and AI across different companies such as BMW Group or Volkswagen Group. He is renowned for his expertise especially in the field of AI in general and for the fields of NLP and Generative AI in particular.

Contributors

The authors would like to thank you the following persons for their invaluable contribution to this publication:

Dr. Paul Yu-Chun Chang

Paul Yu-Chun Chang is an AI Expert specializing in Large Language Models at appliedAI Initiative GmbH. He has 10 years of interdisciplinary research experience in computational linguistics, cognitive neuroscience, and AI, and 5 years of industrial experience in developing AI algorithms in language modeling and image analytics. Paul holds a PhD from LMU Munich, where he integrated NLP and machine learning methods to study brain language cognition.

Simon-Pierre Génot

Simon-Pierre Génot works as a Senior Manager AI at Infineon, where he is responsible for AI strategy and use case development. Previously, Simon worked in machine learning research for IBM Research in the USA before transitioning to the strategic side by launching the first AI initiative at BayWa.

Mingyang Ma

Mingyang Ma works as Senior AI Strategist at the appliedAI Initiative GmbH, supporting all partner companies' decision making and technical solution identification of various AI use cases, with a particular focus on leveraging LLMs. With over 6 years of expertise in NLP, Mingyang has excelled in the realm of Conversational AI, demonstrating her proficiency in application DevOps and platform development across various processes during her tenure at BMW Group in both Germany and the USA.

Dr. Philipp Max Hartmann

Philipp Hartmann serves appliedAI as Director of AI Strategy at the appliedAI Initiative GmbH. Prior to joining appliedAI, he spent four years at McKinsey&Company as a strategy consultant. Philipp holds a PhD from Technical University of Munich where he investigated factors of competitive advantage in Artificial Intelligence.

About appliedAI

The appliedAI Institute for Europe aims to strengthen the European AI ecosystem, develop knowledge around AI, provide trusted AI tools, and create educational and interactive formats around high-quality AI content.

As a non-profit subsidiary of the appliedAI Initiative, the institute was founded in Munich in 2022. The appliedAI Initiative itself is a joint venture of UnternehmerTUM and IPAI. The institute is managed by Dr. Andreas Liebl and Dr. Frauke Goll.

The appliedAI Institute for Europe focuses on the people in Europe. It pursues the vision of shaping a common AI community and providing high-quality content in the age of AI for the entire society. By promoting trustworthy AI, the Institute accelerates the application of this technology and strengthens trust in AI solutions.

With a focus on knowledge development and the provision of trusted AI tools, the appliedAI Institute for Europe provides a valuable resource for companies, organizations, and individuals looking to expand their knowledge and skills in AI. Through educational and interaction formats, the Institute enables an intensive exchange of expertise and fosters collaboration between stakeholders from different fields.

The appliedAI Institute for Europe invites companies, organizations, startups, and AI enthusiasts to benefit from the Institute's diverse offerings and resources.

For more information, please visit www.appliedai-institute.de.

appliedAI
institute
for europe

appliedAI Institute for Europe
Lichtenbergstr. 6
85748 Garching
Germany
www.appliedai-institute.de

Eine gemeinsame
Initiative

UNTER
NEHMER
TUM

lpai