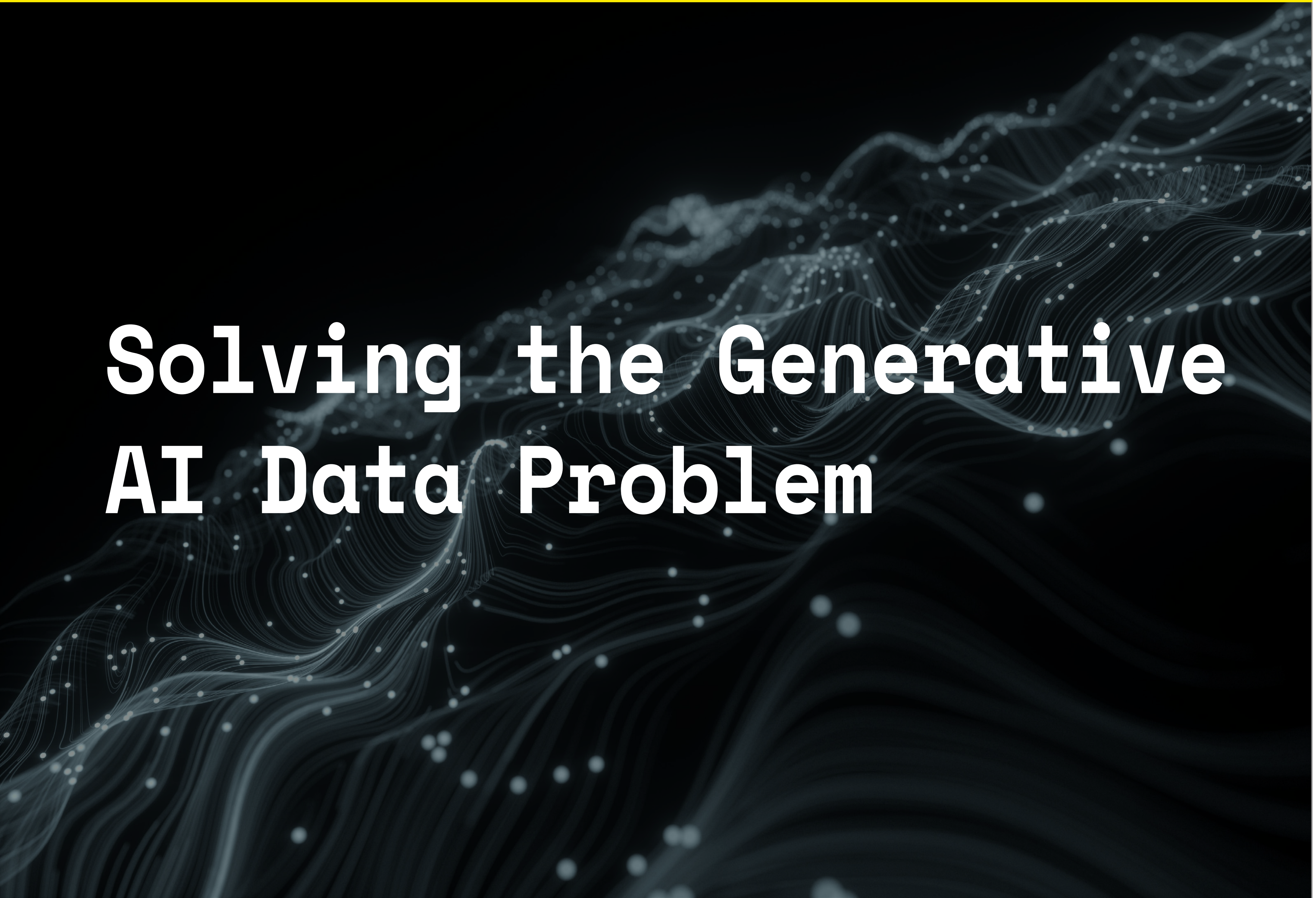
AI BUSINESS

APRIL 2024



Solving the Generative AI Data Problem

INSIDE:

Next

FEATURE

Solving the Generative AI Data Problem

Dealing with structured, numerical and timestamped data is a huge challenge for individual and businesses. AI models urgently need to rise above their limitations and analyse data to create profound, actionable insights. This article considers some of the challenges of current language models.

By: Benjamin Bohman

The progression of artificial intelligence (AI) from small language models (SLMs) to large language models (LLMs) marks significant progress in enhancing usability and accuracy within business environments. However, a considerable obstacle remains; effectively processing structured data, particularly numerical, tabular and timestamped data. This difficulty has emphasized a critical deficiency in the current capabilities of AI, emphasizing the need for innovative solutions that can navigate the complexities of such data with accuracy and understanding. As businesses increasingly rely on data-driven decision-making, the demand for AI models capable of thoroughly analyzing and interpreting structured data has never been more urgent, highlighting a crucial juncture in the effort to maximize the potential of AI in business.

The excitement surrounding SLMs has added a new dimension to AI applications due to their domain-specific efficiency and lower operational costs. However, this enthusiasm is tempered by the realization that SLMs and LLMs as well, have limitations when it comes to handling complex, structured datasets that are fundamental to enterprise operations. This disparity between the potential of generative AI and its practical application in real-world scenarios poses a significant challenge, impacting the functionality of AI solutions and limiting the insights that can be gleaned from rich, structured datasets. As a result, there is a critical need for more capable models to drive the next wave of AI innovation.

The Rise and Limitations of Small Language Models

Small language models have seen a recent rise in AI, noted for their effectiveness and capacity to adapt to specialized topics. Their smaller size, in comparison to LLMs, means they consume less energy and have quicker training times, making them an appealing choice for applications requiring real-time processing. Additionally, the capacity to customize SLMs for specific industries or tasks

Previous Next

has resulted in substantial improvements in productivity and operational efficiency. However, the intricate nature of structured enterprise data often surpasses the processing capabilities of SLMs, revealing a disparity between their potential applications and the depth of analysis needed for nuanced decision-making in business contexts.

This limitation becomes particularly noticeable when SLMs are confronted with data that requires more than just a surface-level comprehension, but rather a thorough, contextual analysis—such as time-series financial records or multidimensional customer data tables. While the specialized nature of SLMs can be advantageous in certain contexts, it also presents a constraint, as their limited understanding and contextual awareness can lead to outputs that lack the necessary relevance for complex decision-making processes. This challenge is further complicated by the need for ongoing model adjustment and the expertise necessary to optimize SLMs for specific tasks, which contributes to the operational complexity and resource demands of implementing AI at a large scale within enterprises.

Challenges in Traditional AI Approaches

The complex process of customizing SLMs underscores a larger issue in AI: finding the right balance between customizability and adaptability. While SLMs are honed for specific tasks, their limitations in handling diverse datasets become evident. This need for specialization requires a high level of expertise in model training and data science, creating a substantial barrier to entry for organizations lacking dedicated AI teams. Relying on specialized knowledge for SLM optimization and upkeep highlights a major obstacle in the widespread implementation and expansion of AI solutions, especially when it comes to using AI for strategic decision-making based on organized data.

The difference in performance between SLMs and LLMs when dealing with structured data brings up concerns about the future path of AI advancement. Although LLMs excel at handling vast amounts of data and producing natural-sounding text, they frequently lack the accuracy needed for tasks reliant on analyzing structured, numerical and tabular data. This disparity underscores the necessity for a new category of models that merge the domain-specific effectiveness of SLMs with the scalability and processing capabilities of LLMs, specifically targeted at the intricate challenges presented by enterprise data.

Trends in Data and Analytics Impacting AI Models

The changing landscape of data and analytics has revealed several trends that add complexity to the challenges experienced by conventional AI models.

Gartner's recognition of value optimization as a primary trend highlights the growing demand for D&A leaders to prove the concrete business results of their efforts. This trend stresses the requirement for AI models that can handle both structured data and convert their analyses into insights that guide important business decisions. Similarly, the trend of overseeing AI risks, such as ethical concerns and data privacy, brings an additional level of complexity to the creation and implementation of AI solutions, requiring models that are not just effective and precise but also transparent and reliable.

The increasing importance of observability and the need for sharing data are notable trends that suggest the emergence of a connected data ecosystem. In this environment, AI models must be able to function smoothly across various data sources and formats. The demand for models that can handle this complexity, offering practical insights while maintaining data integrity and compliance, stresses the limitation of current AI capabilities. These trends depict a swiftly changing digital landscape in which conventional AI methods are insufficient, calling for the creation of models that are flexible, scalable and able to meet the advanced requirements of contemporary enterprise data analysis.

Setting the stage for a new paradigm in AI

The challenges described highlight important turning point in the development of AI. There is a growing demand for a new category of AI models that can effectively handle the diverse nature of numeric data, a task that current models struggle with. There is a noticeable sense of expectancy within the AI community as researchers and developers explore ways to surpass the limitations of existing technologies. The objective is not only to bridge a gap but to redefine the potential of AI in analyzing complex datasets. This effort necessitates a combination

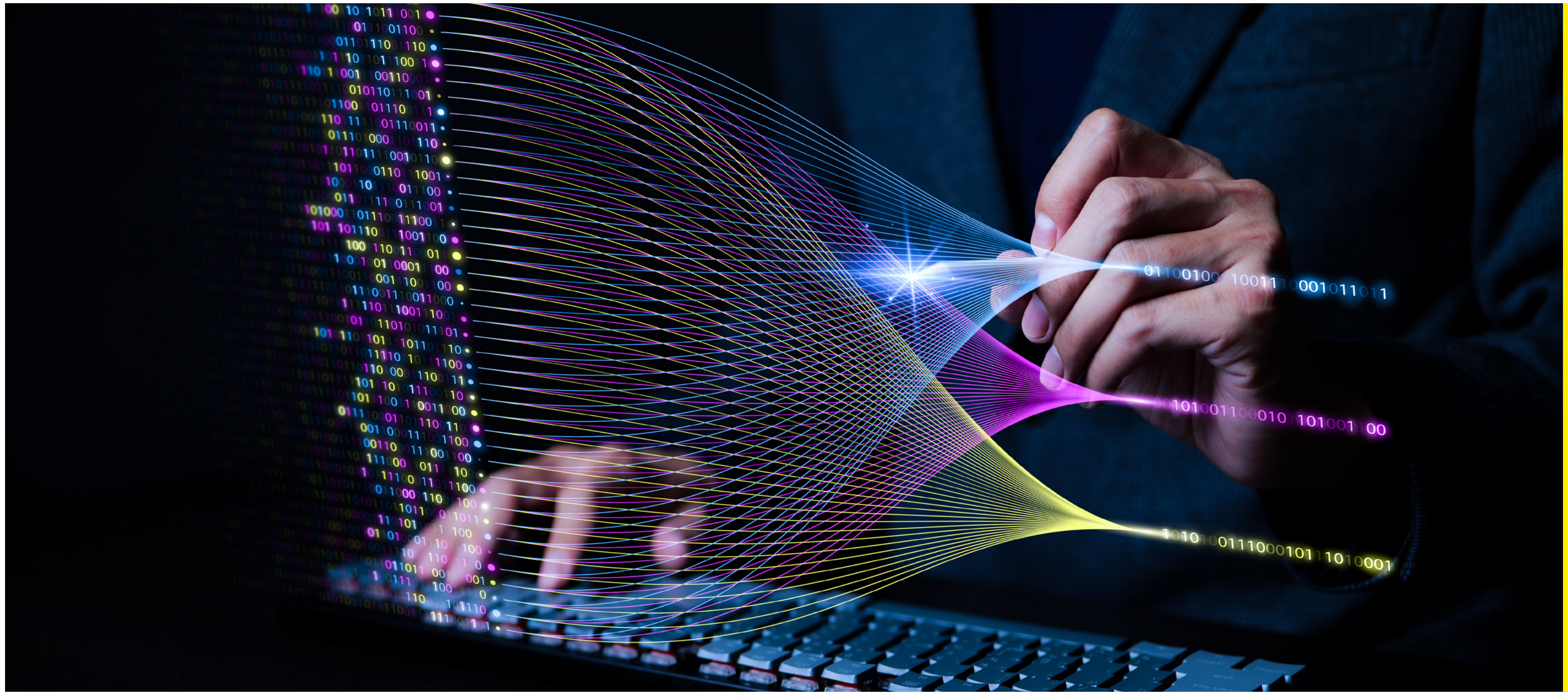
Previous Next

of innovation, precision and a deep understanding of the distinct requirements of structured data. The industry is on the verge of embracing models that can not only dissect and comprehend the intricacies of structured data, but also integrate these insights into practical intelligence, thereby transforming how businesses utilize data for decision-making.

As this story unfolds, the focus shifts to the growing demand for AI systems that combine specialized knowledge with the ability to adapt to diverse data structures. The next wave of AI models aims to bridge this gap by providing robust and flexible solutions that can navigate the complexities of structured data. This evolution in AI marks a significant move towards models that are not only more capable but also more understandable and in line with the strategic goals of businesses. The pursuit of such advanced AI solutions is not just a technical challenge; it signifies a strategic necessity to unleash the full potential of data analytics, ensuring that AI remains a driving force in the competitive business landscape.

Conclusion

The exploration of the generative AI data challenge, particularly when dealing with structured, numerical and timestamped data, sheds light on both the obstacles and the abundant potential that await. This situation emphasizes the urgent requirement for progress in AI models capable of thoroughly examining structured data, offering not only analysis but also profound, actionable insights. As the AI community moves closer to this goal, there is a growing emphasis on creating models that have the required analytical depth and specificity to fulfill the intricate needs of contemporary businesses. These advancements are crucial not only for overcoming current challenges but also for establishing a new standard for the potential of AI in data-driven decision-making.



The shift towards more advanced AI models demonstrates the ever-changing nature of the field, showcasing a wider movement towards innovation and enhancement. The pursuit of creating AI solutions capable of accurately understanding and utilizing structured data is not merely a technological accomplishment; it aims to revolutionize the realms of business intelligence, decision-making and strategic planning. Fundamentally, the quest for these advanced AI models reflects a larger ambition to utilize AI in more practical and relevant ways within the business world. This promises a future where AI can comprehend complex data and enable organizations to derive valuable, actionable insights with unparalleled efficiency and clarity.

NEWS

Gartner: Top 10 Data and Analytics Trends

Value optimization is the top trend identified by Gartner analysts at its data and analytics summit held recently in India.

By Helen Hwang

Data and analytics teams struggle with explaining the value they deliver to the company in a way business leaders understand, according to a new report by Gartner.

As such, data and analytics (D&A) leaders must engage cross-functionally to understand the best way to drive adoption, combining better analysis and insights with an understanding of human psychology and values.

This and other top data and analytics trends were announced at the recent Gartner Data & Analytics Summit in Mumbai, India.

Here are the top 10 trends:

1. **Value optimization:** D&A leaders should clearly convey the connection between business priorities and D&A initiatives by honing their value-management competencies, including value stream analysis, value storytelling, and the measurement of business outcomes.
2. **Managing AI risks:** This includes taking note of new risks such as poisoning of training data, fraud detection circumvention or ethical risks. Creating trust among stakeholders is imperative to the increased adoption of AI.
3. **Observability:** This enables companies to reduce the time it takes to identify the cause of problems that can impact performance and also enables them to tap reliable and accurate data to make timely business decisions. D&A leaders must evaluate their data observability tools in light of the needs of users and how these tools fit into the enterprise overall.
4. **Data sharing is essential:** Whether internally and externally, collaborating to share data is key. "Adopt a data fabric design to enable a single architecture for data sharing across heterogeneous internal and external data sources," said Kevin Gabbard, senior director and analyst at Gartner.
5. **D&A sustainability:** D&A leaders must not only provide analysis and insights of data, but also must

improve their own processes to aid in sustainability. Here, a variety of practices are emerging, which include using renewable energy to power data centers as well as tapping more energy-efficient hardware, use of small data and other machine learning techniques.

6. **Practical data fabric:** A data fabric is a data management design pattern that weaves together all types of metadata to observe, evaluate, and recommend data management solutions. It can generate alerts and recommendations to users and enables business users to confidently absorb the data.
7. **Emergent AI:** With generative AI and ChatGPT rising in popularity, these emergent AI technologies will change how businesses adapt and scale. The next generation of AI technologies will give companies capabilities that are not feasible now.
8. **Converged and composable ecosystems:** These ecosystems design and deploy D&A platforms for seamless integrations, technical interoperability, and governance. With the right architecture, these systems can be more modular, adaptable and flexible to scale dynamically as needed.
9. **Consumers become creators:** The use of pre-defined dashboards will be replaced by dynamic, conversational, and embedded user experiences to address content consumers' needs in real time. Companies can expand the adoption of analytics by providing users easy-to-use automated and embedded insights and conversational experiences so they can be creators.
10. **Humans remain key decision-makers:** D&A leaders are evaluating the role of humans in automated and augmented decision-making since not all decisions can or should be automated. "Efforts to drive decision automation without considering the human role in decisions will result in a data-driven organization without conscience or consistent purpose," said Gareth Herschel, VP analyst at Gartner.

Previous Next

3 Most Common Problems with Small Language Models

Small language models are rising in popularity, but they have problems too. Here’s how to address them

By Tom Taulli

Hugging Face CEO Clem Delangue said this about small language models (SLMs) recently: “My prediction: in 2024, most companies will realize that smaller, cheaper, more specialized models make more sense for 99% of AI use-cases. The current market & usage is fooled by companies sponsoring the cost of training and running big models behind APIs (especially with cloud incentives).”

This is backed up by the momentum in Microsoft’s business. In the latest earnings call, the company announced that customers such as Anker, Ashley, AT&T, EY, and Thomson Reuters were exploring SLMs for generative AI app development. CEO Satya Nadella has noted: “Microsoft loves SLMs.”

This is backed up by the momentum in Microsoft’s business. In the latest earnings call, the company announced that customers such as Anker, Ashley, AT&T, EY, and Thomson Reuters were exploring SLMs for generative AI app development. CEO Satya Nadella has noted: “Microsoft loves SLMs.”

Why the excitement? SLMs, which are generally five to 10 times smaller than large language models (LLMs), offer compelling benefits.

“They use less energy and have lower latency,” said Sudhakar Muddu, who is the CEO and cofounder of Aisera. “The training and inference times are also quicker. And the small size means you can use an SLM on the edge. But the most important benefit for the enterprise is that they can be tailored for certain domains and industries. This is where you get the gains in productivity.”

However, he does point out that there are challenges with SLMs. The technology is still in the nascent stages andis complex.

Here’s a look at some of the most common issues and what can be done about them.

#1 - Performance

SLMs are closing the gap with the capabilities of LLMs, in areas such as accuracy. But the differences can still be noticeable and result in a lower-performing application.

"Their limited understanding and contextual awareness often mean they struggle with complex or niche topics, leading to responses that may not be as relevant or coherent as those generated by larger models," said David Guarrera, a principal with EY Americas Technology Consulting. "This limitation impacts not just the depth of knowledge these models can access but also their ability to maintain context over longer interactions."

This is why there should be due diligence about the tradeoffs between SLMs and LLMs. The performance of an SLM can also be significantly improved with fine tuning. In other words, SLMs often do not make much sense when used out-of-the-box.

#2 - Expertise

A common way to optimize an SLM is to use retrieval-augmented generation (RAG). This involves using semantic search – such as with vector databases – to process relevant data. This can improve the accuracy of the generated content as well as allow for more updated results.

But building beyond RAG requires someone with a deeper understanding of AI – and this talent is in short supply.

"The next step in complexity is fine-tuning a model, in which you take an existing AI model and introduce new training data to hone it to a specific data set," said Hymel. "This is more complex because it requires custom data curation, tagging, and running the training, which goes beyond typically generic backend engineer skill sets."

Enterprise generative AI applications may also involve multiple SLMs, which adds to the complexity. For example, there will be a need to work with orchestration tools, such as Kubernetes.

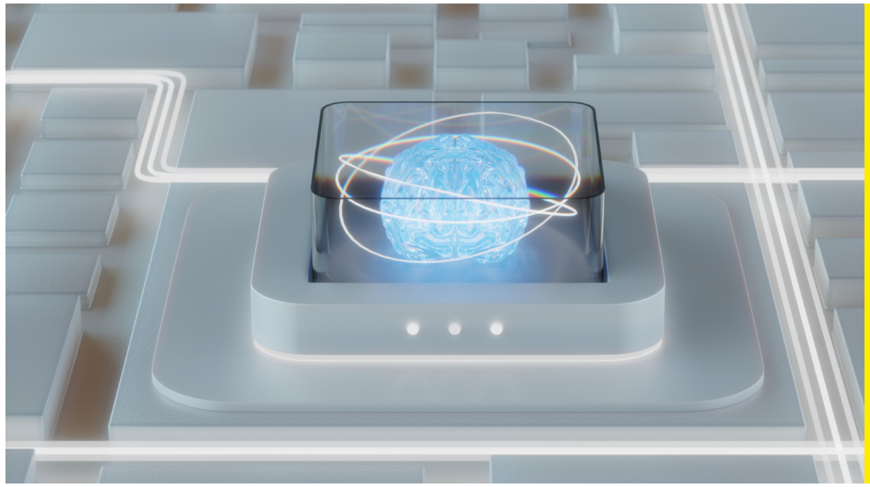
"As you can imagine, training a single model is one thing, but looking to train multiple models to work together is much more difficult. With this more complex architecture, you can increase total cost of ownership, time to market, and initial upfront investment," Hymel said.

"Similar to finding AI talent in general, finding the subset of that group with experience in building these types of systems is even more slim. We expect that the large model providers such as Microsoft, OpenAI, and others will begin offering 'orchestration as a service' to help gain larger market adoption."

#3 - Security

Many SLMs are open source. This allows for more control over the security. For example, an enterprise can deploy an SLM in an on-premise environment.

However, there are still notable issues. "The foremost security risk when using a fine-tuned SLM is data theft and privacy concerns," Mehrin Kiani, who is an ML scientist at Protect AI. "This is especially prevalent if an SLM is fine-tuned on proprietary and confidential data."



"Training models on adversarial examples and implementing detection mechanisms can help identify and mitigate malicious inputs," said Tal Furman, who is the director of data science and deep learning at Deep Instinct. "Other best practices are to implement strong access controls, logging, and monitoring for open-source models."

As with any software that handles sensitive information, there should be robust security reviews for every step of the fine-tuning and operationalization of the SLM.

However, "it is important to note that no security measure can guarantee complete and robust security of SLM-based applications," said Kiani. "The security posture of these can be improved by designing with security-first principles. An insecure GenAI application is useless no matter how unique and wonderful it is."

NEWS

Survey: Data Decision-Makers Find AI Explainability ‘Challenging’

Capital One, Forrester report say 3rd party partnerships will drive ML maturity.

By Ben Wodecki

Companies are embracing AI but they are encountering problems in deployment, with nearly three-quarters of data management decision-makers citing issues with model explainability, according to a survey from Capital One and Forrester.

The report said 73% of respondents faced difficulties in transparency, traceability and explainability of data flows when they try to deploy and scale machine learning to more use cases.

“Businesses see massive potential in applying machine learning but encounter headwinds in their data,” said Dave Kang, senior vice president and head of Data Insights at Capital One. “This can hinder businesses from seeing actionable insights, and perversely shy away from adopting and operationalizing ML solutions in the first place.”

Another key obstacle was breaking down data silos, with 57% believing internal silos between data scientists and practitioners inhibit ML deployments. Also, 38% said that data silos across the organization and external data partners pose a challenge to ML maturity.

Diverse, messy data sets (36%), difficulty translating academic models into deployable products (39%) and reducing AI risk (38%) were pain points for data decision-makers as well.

“To overcome challenges, organizations must focus on the business outcomes of ML and build partnerships with proven leaders in their ML journey,” according to the report.

Moreover, “without better explainability and transparency” top executives and directors “have trouble seeing business benefits after adopting AI/ ML solutions.”

“If there’s no clear connection to ROI, executive buy-in decreases, which reinforces data silos, creates struggles in driving actionable insights and inhibits operationalization,” the report warned.

Capital One and Forrester surveyed 150 North American data decision-makers to determine their ML goals and challenges.

Partnerships to drive ML maturity

According to the survey, 67% said they intend to leverage partnerships to fill ML staff gaps. Around 37% said they’re currently partnered with a third party for ML model development and plan to grow that collaboration.

Close to one-fifth of respondents’ organizations plan to begin a partnership in the next year. Just 11% said they weren’t partnered with an outside group but were interested in potentially joining one.

“To push their organizations out of the experimentation phase, decision-makers should seek out ML partners that have firsthand experience in building and operationalizing ML applications,” the report said.

“These partnerships will help organizations resolve their explainability, transparency, and skills gap issues and create an AI/ML ecosystem and community of practice, setting organizations up for success in scaling their ML strategies.”

But overall, concentrating on business outcomes for ML work is the key strategy that will “drive ML maturity,” the report concluded.

Previous Next

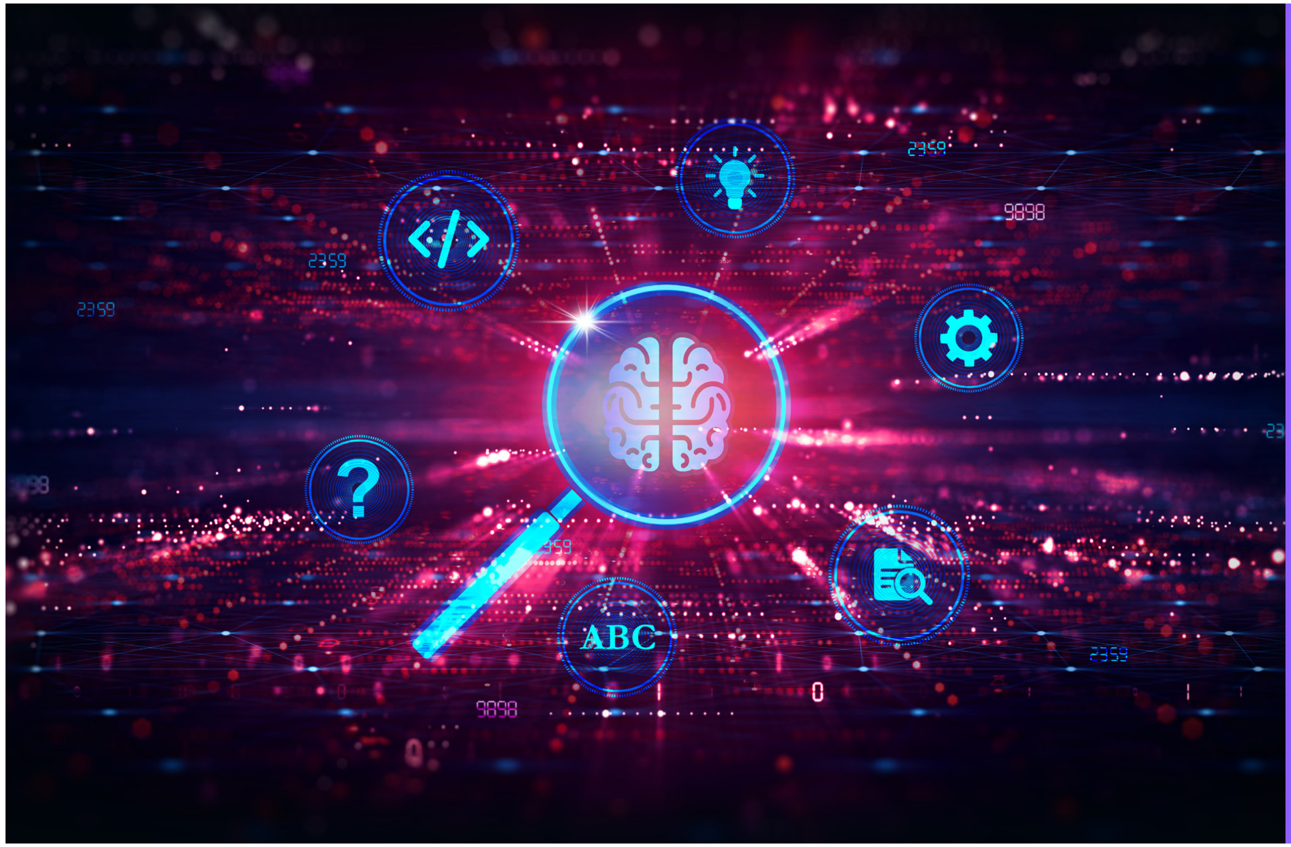
LLMs are Not a Panacea: Challenges, Concerns, and Shortcomings of the Technology

Large language models have captured the imagination of many, but they have computational limits when applied to complex, time-series datasets. Large graphical models have emerged to address this gap, helping organizations generate insights, predict business outcomes, and plan for intricate business environments.

By Gopkiran Rao

Generative AI has taken the world by storm, led by the buzz around large language models (LLMs) such as ChatGPT from OpenAI. In this environment, it's not surprising that the AI-curious have come to equate generative AI with LLMs. But there is another significant approach that must be considered – large graphical models (LGMs). Unlike LLMs, LGMs are purpose-built for harmonizing and building predictive insights from tabular, time-series data. In a world of infinite possibilities, LGMs are transformative for helping businesses identify the best path forward, aligned to desired business outcomes; in short, they are the future of forecasting and planning.

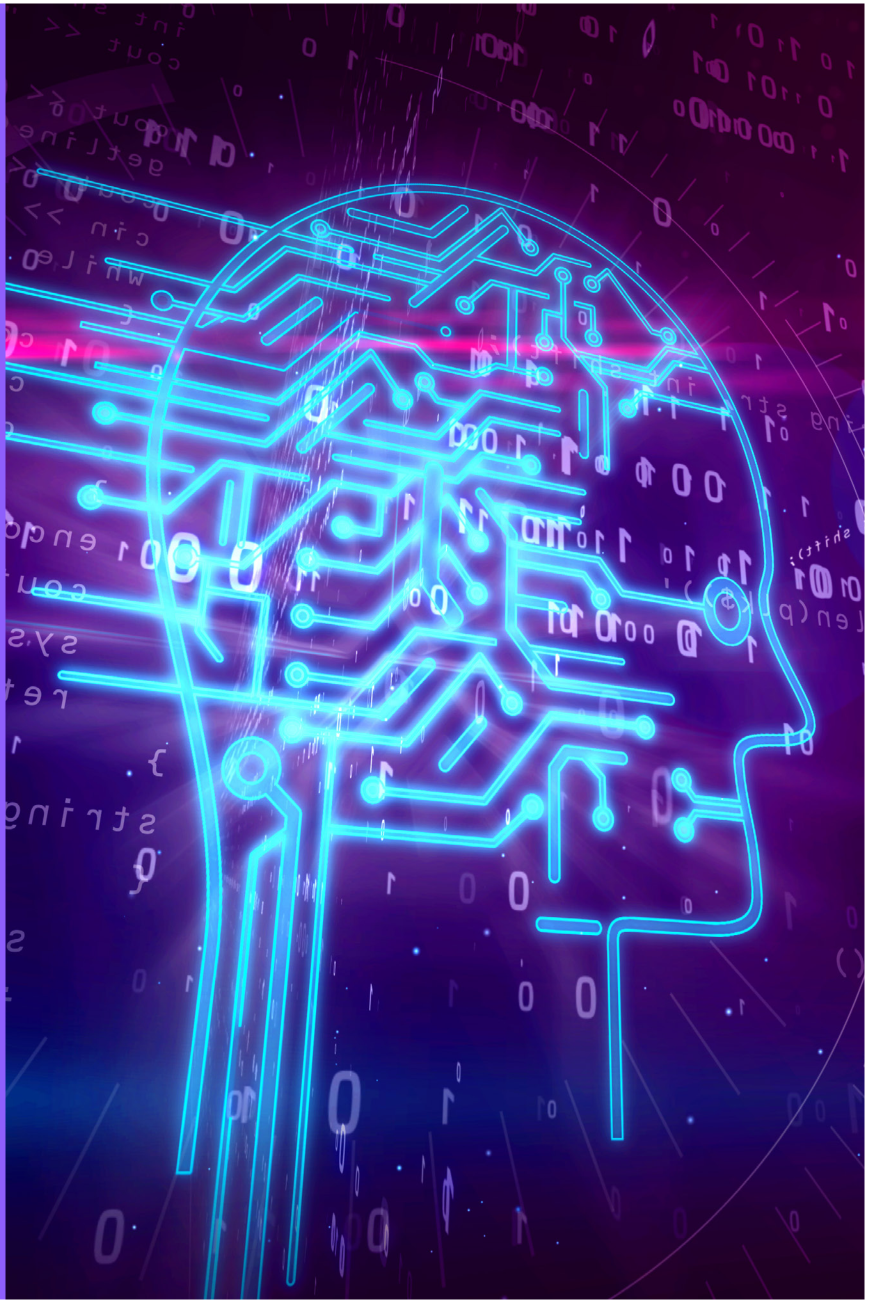
It's understandable that LLMs have captured our attention so quickly. They've demonstrated significant business value for certain text and image-based use cases such as content development, research, and translation. However, LLMs come with challenges, such as hallucinations, contextual relevancy, and legal risks due to use of proprietary information. Their inherent qualities make LLMs both unreliable and inefficient for use cases like forecasting, which require the synthesis and analysis of numbers in a specific business context.



LLMs are costly and come with risks

One of the biggest challenges of running LLMs at scale is the cost. One analysis

Previous Next



found that for a large organization using GPT-4 to analyze longer documents and ask a million questions per day, yearly costs would range between $28.47 million and $56.94 million per year. More and more organizations have LLM projects stuck in pilot mode because the ROI doesn't justify the investment required to deploy at scale. According to a Gartner survey from late 2023, 55% of organizations are experimenting with generative AI, but only 10% of organizations actually have something in production.

Data requirements

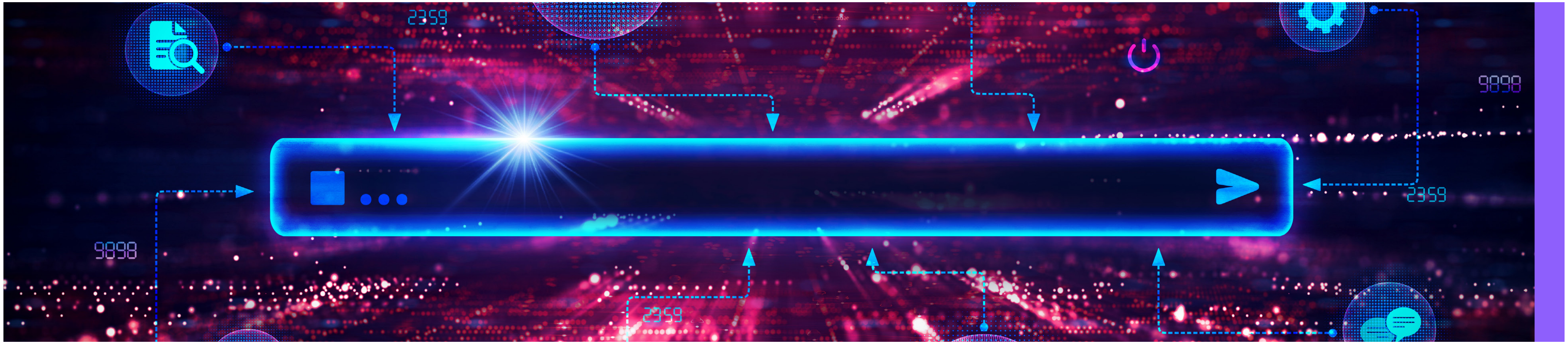
One reason LLMs are so expensive is that to provide truly relevant content they must be trained on the lexicon of the specific business environment they are working within. This may require massive volumes of data to effectively learn and provide accurate answers. From an infrastructure perspective, collecting and analyzing all this data results in high computing and storage fees, and takes time. Consider that LLMs like ChatGPT had to be trained on the entirety of the web before they were released to the public, costing hundreds of millions of dollars. As Forbes noted, "When asked at an MIT event in July whether the cost of training foundation models was on the order of $50 million to $100 million, OpenAI's cofounder Sam Altman answered that it was 'more than that; and is getting more expensive."

Training costs

In addition to the startup costs for employing the foundational model platform, organizations must additionally invest in training the models for their specific use, which requires massive amounts of data, and further investments in data cleanliness to ensure the right data is being used for training. Using large language models (LLMs) without incorporating current, accurate data tailored to a specific business and industry, like the latest inventory levels or marketing promotions, can lead to disastrous outcomes for organizations. Major players are addressing these challenges with more efficient algorithms, code optimization, and cheaper GPUs. However, there's still a long way to go – likely several years – until LLMs achieve true affordability for most enterprises. In the meantime, it doesn't matter what sort of exciting things LLMs can do: If they don't provide economic value, businesses won't deploy them in production.

Privacy challenges

Another challenge with LLMs is data privacy and compliance. A survey from Malwarebytes found that 81% of respondents were worried about the security and safety risks of ChatGPT. There are many concerns here. From an enterprise perspective, the biggest issue is that organizations have no way of knowing for certain if the data they're sharing with LLMs is being kept safe and secure. There are additional liabilities as well. For instance, if an



enterprise's LLM is leveraging copyrighted materials, they could face lawsuits down the road if they create products or services that were informed by those materials. With The New York Times' lawsuit against OpenAI ongoing, this issue may not be resolved for a few years.

LLMs fail to predict business outcomes via numerical data

The limitations of LLMs go beyond the challenges highlighted above. Even once LLMs become cheaper and safer, there will still be generative AI tasks they're just not suited for. Lots of businesses think they can use LLMs to assess or predict certain outcomes and events, but LLMs weren't designed for activities like data reconciliation, forecasting or planning, all of which require making connections between seemingly unrelated tabular time-series datasets and variables.

LLMs are designed to synthesize, recognize, and reassemble mostly unstructured text data (i.e. web content and social media posts) for use cases such as creating content and translating languages. Unless explicitly trained on pre-sorted, curated content, they aren't able to effectively ingest, collect and analyze company- and context-specific structured data such as inventory, sales numbers, tax payments, sensor data, statistical data, IT event data and more. This renders LLMs completely ineffective for describing and modeling business trends.

Here's an example of where an LLM falls short. A large retailer (with a considerable online presence) wants to optimize planning operations and supply chain logistics heading into the busy holiday season. They're looking for answers to questions such as, "How many seasonal workers should we hire?" and "How many days should we expect it to take to ship packages to customers?" To get those insights, the LLM would need to analyze things like past sales data, delivery times, lead times, labor rates, weather data, postal shipment data, and more. But as LLMs aren't trained on that data; or designed for this kind of analysis, the model might provide broad guidance based on secondary sources such as industry case studies, media articles and or published reports about major holiday work and supply chain trends. To put it simply, when precise, accurate, and relevant answers within confidence intervals are required, LLMs just can't do the job.

LGMs fill the gap

An LGM (large graphical model) is a model that can visualize the complex relationship between different data points, and then be used to determine how those relationships may change in the future as the data changes. Probabilistic in nature, LGMs use historical data as well as the influence of other variables to help businesses predict future outcomes, providing a range of future possibilities as well as the confidence level of those possibilities.

LGMs differ from LLMs in several key ways:

- **Modelling tabular, time-series data:** Unlike LLMs, LGMs are designed to learn the model of tabular, time-series data without any teaching or supervision. LGMs were created specifically for the purpose of forecasting and incorporating assumptions, making them ideal for use cases like the retail example above, as well as for use in finance, healthcare, and other sectors with lots of tabular, time-series data.
- **Small data, low costs:** LGMs require less data than LLMs and therefore have lower compute and storage requirements, driving down costs. This also means that organizations can get accurate insights from LGMs even with limited training data.
- **Safety:** LGMs support better data privacy and security. They train only on an enterprise's own data with supplementation from select external data sources (such as weather data and social media data) when needed. There is never a risk of sensitive data being shared with a public model

Conclusion

With safeguards, LLMs have shown benefits for individuals and even business—especially when it comes to generating data that is not mission critical to the planning or optimization of key resources. But there are some hurdles that need to be overcome before they'll be computationally efficient and trustworthy enough to truly deliver on their unsupervised value. And even when those challenges are solved, there will still be tons of use cases that are unserved by LLMs and will never be served by LLMs, as the technology simply wasn't built for these purposes.

LGMs were created specifically to work on tabular, time-series data to build insights, predict outcomes, and plan for complex environments. Moreover, LGMs have already solved the problems plaguing LLMs and are ready for full deployment today. The technology is cost effective, secure, requires minimal training data, and only needs an organization's own data to work. Ikigai is the industry pioneer in large graphical models. Visit here to learn more about how Ikigai's LGMs enable organizations to see into the future and make better business decisions today.