Persistent

The Right Way Forward with Data for Generative AI:

Developing Winning Strategies for Business Value

Introduction

A recent survey from Boston Consulting Group found that 89% of executives rank AI and GenAI as a top-three tech priority for 2024. Enterprises see the potential to bring new process efficiencies, improve employee productivity, and deliver business value. The ability of any business to use and extract value from GenAI is heavily dependent upon a number of essential elements. This white paper details Persistent's knowledge and point of view around those elements to develop and execute GenAI initiatives.

Virtually all enterprises today are investigating how Generative AI (GenAI) can bring new process efficiencies, improve employee productivity, and deliver business value. The readiness of any business to leverage GenAI is heavily dependent upon possessing the right data strategy. Without one that's well-defined, a business may struggle to collect, store, and manage the data needed to either use third-party GenAI models or fine-tune foundational models, or train GenAI models from scratch. Additionally, a poorly designed data strategy may result in biased or incomplete data, which can lead to inaccurate or unreliable AI-generated outputs. Therefore, a business must have a clear understanding of its data needs and develop a comprehensive data strategy to ensure that it can effectively leverage GenAI to drive innovation and competitive advantage.

Achieving a successful data strategy, and subsequent success with GenAI, requires an organization to focus on identifying the right GenAI use cases, what data is required to power the large language models behind GenAI, and what behavioral, technical, staffing and culture requirements are needed. In this white paper, we'll explore a path forward that enterprises can follow to attain business value and growth by leveraging data in GenAI.

Strategy Comes First

To chart a successful path the maximizes your data assets for GenAI, enterprises should resist the urge to rush into one-off use cases or make investments that are not tied to a well thought-out strategy. In other words, strategy must come first, followed by a focus on data quality, security, and testing, to determine how to properly leverage data for GenAI use cases. Let's first review what strategic elements should be addressed before moving into how to operationalize a data strategy for GenAI.

Identify Generative AI Needs

Counter to the belief that GenAI is a panacea for all analytical problems, there are many occasions where the use case at hand can be better dealt with the use of Machine Learning or Statistical Modeling rather than applying GenAI. So, it's important to consult experts and identify the areas where GenAI can bring real business value, versus a use case where other technologies would suffice.

As an enterprise plots out a strategy, viewing use cases through a lens of GenAI complexity provides a better view of data implications.

Low Complexity: These use cases would require the application of preexisting foundational or large language models (LLMs) through UI or API calls. Here a team needs to make sure that quality data is available for forming GenAI-usable prompts as input to models and model output is validated to align with business values. Data security is another important aspect to make sure that no sensitive data exfiltration happens due to an external API call.

Medium Complexity: These use cases require fine-tuning of foundational models so that it can better serve specific use cases in hand. Here a team needs to derive strategy around creating and managing training data required for fine-tuning.

High Complexity: These are either very specific industry use cases or use cases which deal with super-sensitive customer data. In these cases, foundational models need to be built from scratch. Here correct strategies around training, data collection, storage, data labeling, data architecture, and data platform need to be adopted.

Implement the Right Data Architecture

GenAI demands data modalities beyond structured data. It has expanded the scope of what would be considered valuable data to include unstructured data like chats, code, reviews, reports, images, audios, and videos. This is a significant shift as data organizations have traditionally only worked with structured data in tables. To take advantage of this shift, organizations need to focus on right data architecture approach. Existing data architectures need to be updated, and here are some key data architecture components for a GenAI setup.

Data Lakes

Data lakes provide a scalable and flexible foundation for storing and managing diverse data types. They are designed to quickly ingest data of any type or size. This data can be used to create model training data. Some popular data lakes include Azure Data Lake Storage, Amazon S3, Databricks Lakehouse Platform, Snowflake Data Lake, etc.

Data Puddles

Data puddles are proprietary data sets that are curated from the wider and deeper data lakes. These are small datasets that are usually built for a specific use case or business objectives. These can be pre-built by an engineering team or created on request. Data puddles can help to provide data for model fine-tuning (which we review in subsequent section), where fine-tuning is performed to improve foundational models on a specific business case.

Data Products and Data Mesh

Data products and data mesh are two concepts that can be used in GenAI to streamline the data acquisition process and improve model performance. Data products are data organized in a format that can be easily consumed, ensuring that GenAI models are supplied with high-quality and relevant data. Data mesh is a decentralized approach to data architecture and management that places the responsibility of data management and governance in the hands of the teams that actually use the data. This approach breaks away from traditional siloed and centralized data handling methods, reducing bottlenecks, and increasing efficiency. By leveraging data products and data mesh, organizations can ensure they have access to quality, timely, and relevant data, transforming the cumbersome data acquisition process into a smooth, manageable, and efficient system that paves the way for businesses to harness the power of GenAI.

Vector Databases

Vector databases store and provide access to unstructured data, such as text or images, in the form of their vector embeddings, which are the numerical representation of the data as a long list of numbers that captures the original data object's semantic meaning. Because similar objects are close together in vector space, the similarity of data objects can be calculated based on the distance between the data object's vector embeddings. To completely leverage GenAI functionalities, support of vector databases is critical, and some of the more widely used ones include Chroma, Pinecone, Vespa, and Qdrant.

Graph Databases

A graph database is a one that uses graph structures to store and represent data. Graph databases are useful for data with complex relationships because they can quickly locate connections between points in the graph. Graph databases use nodes, edges, and properties to represent data. Nodes and edges represent data items in the store, and edges represent the relationships between the nodes. Graph databases can be used to ground LLM responses in validated facts. Some of the popular graph databases include Neo4j, ArangoDB, and Amazon Neptune.

Install a Data-Centric Culture

A data-centric culture can be thought as the set of shared beliefs and behaviors of people in an organization who value, practice, and encourage the use of data to improve decision-making — and creating such a culture is a shared responsibility for everyone.

Key enablers for data-centric culture include:

Leadership commitment: Leaders who champion the use of data and set a good example for their teams.

Data accessibility: Data should be accessible by employees and stakeholders with the right control mechanism in place.

Data literacy: A high data literacy level empowers an organization to ask the right questions, acquire pertinent data, derive insights, validate assumptions, and make decisions.

Data-driven decision making: All business decisions, whenever relevant, should be based upon data-driven approach in order to be relevant and timely.

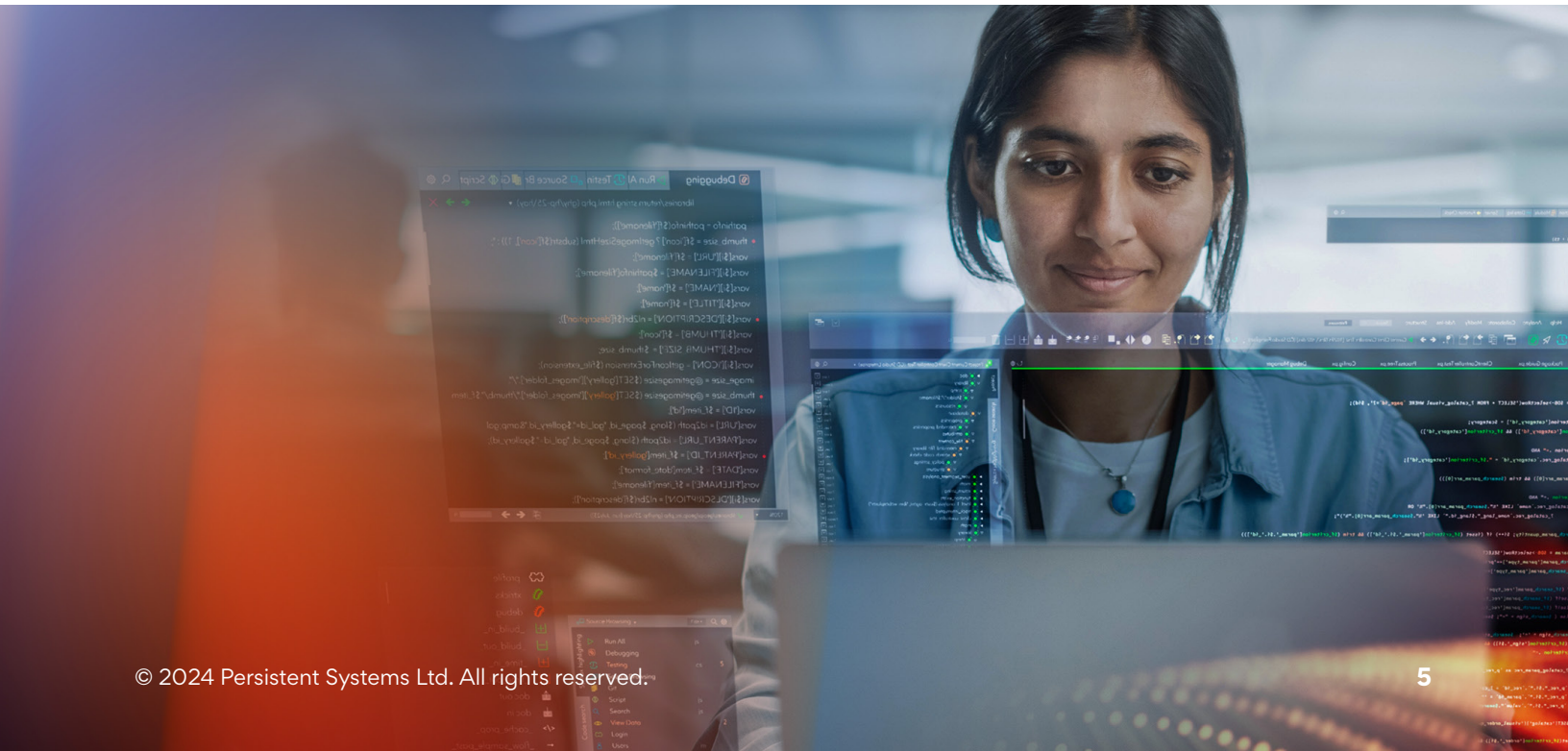
Manage the Data Lifecycle

Data lifecycle management is critical for GenAI to enable proper training of LLMs so GenAI can generate new content and make predictions. Proper data management ensures that data is accurate, relevant, and up to date, and if data is not properly managed, it can lead to inaccurate or biased results, which can have serious consequences in fields such as healthcare or finance.
The data life cycle management process involves several stages, including data collection, cleaning, labeling, storage, and analysis — and each plays an important function. For example, data cleaning and labeling can help remove data errors and inconsistencies, while data analysis can help identify patterns and trends that can be used to improve the accuracy of the AI models.

Manage Talent

Effective talent management for GenAI involves identifying and recruiting individuals with the necessary skills and experience, providing ongoing training and development opportunities, and creating a supportive and collaborative work environment. It also involves ensuring that team members have access to the latest tools and technologies and can stay up to date with the latest developments in the field. In addition, the application of GenAI use cases within an enterprise has the potential to open up new opportunities for data scientists, analysts, and others that can be tasked with identifying new use cases that could drive growth and value throughout an organization, given their experiences with GenAI and how models perform.

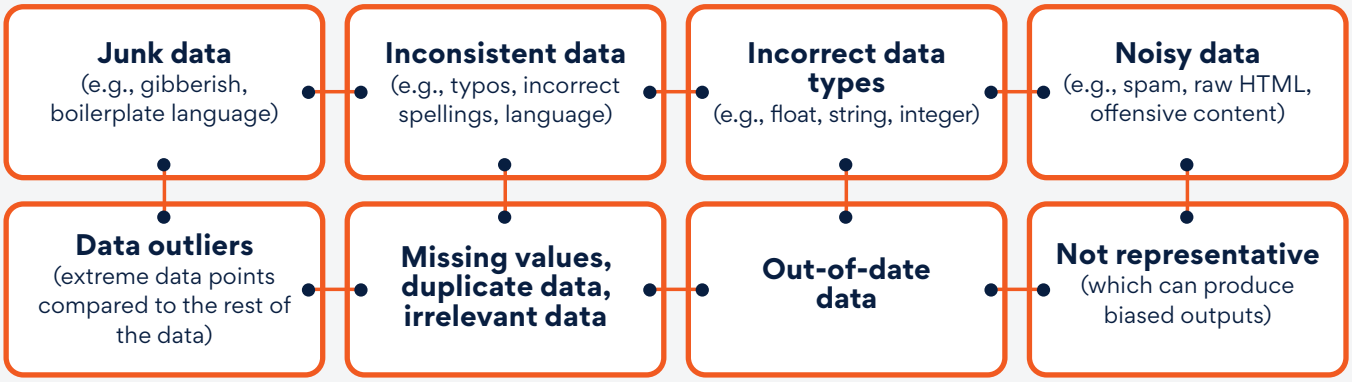
Graph databases can be used to ground LLM responses in validated facts. Some of the popular graph databases include Neo4j, ArangoDB, and Amazon Neptune.



With a Strategy in Place, Focus on Data Quality for Generative AI

The criticality of data quality to develop models for GenAI cannot be overstated. High-quality data ensures that the models learn from reliable and representative examples, leading to the generation of high-quality and relevant content. Therefore, data cleaning is an important part of modern dataset creation for model training. Poor data quality, meanwhile, can lead to models that generate irrational or biased outputs, and could ultimately doom any GenAI strategy (see Figure 1).

Figure 1: Data issues which can lead to poor data quality



Improving data quality can dramatically change the shape of the scaling laws, which for neural language models describe how model performance can improve with different scaling factors including size of the training data and number of model parameters. It is also crucial to recognize the significant impact that poor data quality can have on the training process. When data is of low quality, it can introduce noise, biases, and inaccuracies, leading to a range of undesirable consequences, including:

Biased Outputs: GenAI models may generate text that reflects the biases present in the training data, perpetuating harmful stereotypes and discrimination.

Inaccurate information: If the data contains inaccuracies, the model's responses may be factually incorrect, leading to misinformation. This can be particularly problematic in fields such as healthcare, where inaccurate information can have serious consequences for patients, or banking where decisions can impact personal finances.

Unintended consequences: Such errors can lead to contextually inappropriate or irrational outputs, which can be confusing or misleading for users, also referred as model hallucination.
Furthermore, smaller models with quality data requiring less training can significantly reduce the environmental cost of LLMs – it takes scalable compute and storage to power LLMs for GenAI, so more efficient data leads to more efficient energy output in cloud/data centers.

Furthermore, smaller models with quality data requiring less training can significantly reduce the environmental cost of LLMs – it takes scalable compute and storage to power LLMs for GenAI, so more efficient data leads to more efficient energy output in cloud/data centers.

Types of Data Utilized in Generative AI

GenAI data primarily refers to the information used to train the models. This data can include text, images, audio, or video depending upon the type of the model. Models learn patterns from this data, enabling them to generate new content matching the input data's complexity, style, and structure. GenAI language models are trained using unstructured and semi-structured data:

Unstructured data refers to any data that does not have a predefined structure or organization. This type of data can come from a variety of sources, such as social media posts, emails, or audio recordings.

Semi-structured data is a type of data that has some structure but is not fully organized. This type of data is typically found in documents such as XML or JSON files, where there is a defined structure but the data within each element may vary.

To build foundational models, which mostly focuses on learning language constructs, no data labeling is required, and these models primarily rely upon open source datasets (see Figure 2). However, to fine-tune the model for a specific task, labeling of the training data in accordance with the business case objective is very much essential.

Figure 2: Open source datasets are instrumental in enhancing the generalization capability of LLMs across a broader context. Here are some popular open source datasets used for training LLMs:

Common Crawl: This dataset comprises terabytes of raw web data extracted from billions of web pages. Several large language models, including GPT-3, LLaMA, OpenLLaMa, and T5, were trained with Common Crawl.

RefinedWeb: RefinedWeb is a massive corpus of deduplicated and filtered tokens from the Common Crawl dataset. It has more than 5 trillion tokens of textual data, of which 600 billion are made publicly available. It was developed as an initiative to train the Falcon-40B model with smaller-sized but high-quality datasets.

BookCorpus: BookCorpus turned scraped data of 11,000 unpublished books into a 985 million-word dataset. The dataset was used for training LLMs like RoBERTa, XLNET, and T5.

The Pile: The Pile is an 800 GB corpus curated from 22 diverse datasets, mostly from academic or professional sources. It enhances a model's generalization capability across a broader context and was instrumental in training various LLMs, including GPT-Neo, LLaMA, and OPT.

C4:C4 is a 750 GB English corpus derived from the Common Crawl. It uses heuristic methods to extract only natural language data while removing all gibberish text. C4 has also undergone heavy deduplication to improve its quality. Language models like MPT-7B and T5 are pre-trained with C4.

Data Pre-Processing Techniques for Generative AI

Data pre-processing is vital for creating training data for GenAI models, as the quality of the training data has a direct impact on model performance. Data pre-processing involves cleaning, transforming, and organizing data to make it suitable for model use (see Figure 3).

Figure 3: The key steps involved in data pre-processing for GenAI models:

Data Cleaning: Removes any irrelevant or duplicate data. This includes removing any special characters, punctuation marks, and stop words. This step ensures that the data is free from any noise or unnecessary information.

Tokenization: Breaks down text into individual words or tokens to make the data more manageable and to enable the LLM model to process it more efficiently.

Text Normalization: Converts text to a standard format. This includes converting all the text to lowercase, removing any accents, and expanding any abbreviations.

Stop Word Removal: Removes common words that do not add any value to the text to reduce the size of the data and to improve the accuracy of the LLM model.

Stemming or Lemmatization: Reduces the words to their root form to reduce the size of the data and to ensure that the LLM model can recognize different forms of the same word.

Encoding: Converts text data into a numerical format that can be used by the LLM model, using techniques such as word embedding.

Challenges in Collecting and Preparing Data for Generative AI

"If people knew how hard I worked to achieve my mastery, it wouldn't seem so wonderful after all."

- Michelangelo

The above quote, where Michaelangelo references the years that he dedicated to his famous works of art, is fairly applicable to GenAI as well. Behind all the magic and wonder that has allowed famous GenAI models such as ChatGPT or DALL-E is immense amount of hard work to help these generative models achieve various and growing levels of mastery.

For example, ChatGPT is a collection 175 billion numbers learned through a highly computationally intensive process by feeding data worth 0.5 trillion tokens or ~1.25 billion pages. The updated version known as GPT 4.0 is said to have trained on 3.5 trillion tokens (7x more!). So, it's not surprising that many enterprises encounter challenges when creating their own GenAI models – from scale and engineering issues to data quality problems and safety concerns.

It's a task to meet those data-related challenges on a smaller scale but when faced with the scale typically required of GenAI models, the challenge becomes exponentially more difficult. And one doesn't have the benefit or luxury of manually checking all data (which if done right is probably the safest way).

Thanks to emerging cloud technologies, big data handling and processing is relatively easier to handle. But for data quality and safety, with manual audit virtually impossible, one is left to use AI itself to assist in filtration of data, redacting specific elements, etc. However, in a world of AI models or heuristic rules, it is not realistic to expect total assurance that there will be no errors. Not surprisingly, it's a challenge that even the most tech-savvy companies and IT leaders are consistently trying to remediate.

Data Augmentation Techniques for Generative AI Inference

After all the efforts that go into achieving GenAI "magic," unless you have specifically trained your own model, the model has not learned from a specific dataset. For example, your company may have a bulky collection of compliance policy documents, internal knowledge bases etc., which were not used in the training phase.

And even if it happens that these documents were used in the model, it is not a guarantee that every bit of information can be retrieved by a GenAI model when it creates an output. Therefore, it's critical to guide the model into generating the right answer by augmenting its input with some relevant data. This data must in turn be retrieved from a suitable data store.

This technique is called Retrieval Augmented Generation (RAG). In summary, RAG allows models to access a specific set of data points which allow it to answer questions more accurately. This is especially important when the information source wasn't used in the training phase.

Most available RAG implementations have focused on querying text sources — structured or unstructured — to answer natural language queries. A key initial step in enabling this on unstructured documents such as PDF files, images or scans is the process of data extraction. While many extraction tools exist, they are limited to the simplest of tables. Some tools claim to be able to handle more complex tables, however the most complex tables, charts and graphs are difficult to extract in any reliable manner in order to enable RAG workflows atop them.

To handle this complexity, it is becoming clearer that the future will include what's being called "multi-modal RAG." In this approach, pages or images are directly converted into an embedding, and an attempt is made to produce a natural language summary of the same. At query time, relevant images are passed along with the query to a multi-modal model. The model then reasons out the answer from the image using the latest state-of-the-art models such as GPT-4 Vision, LLaVA, Gemini, and others.

Considerations for Fine-Tuning

While fine-tuning is not as complex and expensive as training an LLM from scratch, it is still a substantial operation and must be carefully planned. While there are many fine-tuning tools available using a variety of algorithms, the biggest piece still remains the data.

A common problem reported among practitioners is that responses they get from an LLM become worse after fine-tuning. While inappropriate use of fine-tuning algorithms is one cause, the more common problem is around data. The data must be carefully crafted and curated for fine-tuning to deliver its results. In general, a recommended approach is that through simple prompting, and techniques such as few-shot learning, the limitations of the LLM must be fleshed out. This helps identify the Achilles heel if any for that model. Using this as a guide, a careful dataset must be assembled which has a high chance of enabling the LLM to perform better.

It is worth highlighting a common misconception that fine-tuning and RAG are mutually exclusive — which is incorrect. RAG enables an LLM to have access to a specific data fragment which is likely to contain the answer to a user query. So, in any scenario where data keeps changing and query responses need access to this data, some form of RAG implementation, rather than fine-tuning is needed.

Data Security and Privacy Concerns in Generative AI

One very crucial step in preparing data is to scrub private and/or sensitive information from the data being fed into the model — which given the sheer size of data involved is usually easier said than done. In general, there is no single solution for any kind of security that simply works, however enterprises can compensate by introducing additional layers of security for GenAI:

Engineering prompts for safety: Prompt engineering refers to the way inputs are provided to the model so as to produce the most desirable output. The inputs comprise of system-level instructions which govern the behavior of the model (e.g., use concise language, assume the persona of a consultant or teacher, etc.), the actual query or question, and any context information (such as with RAG). However, if not done correctly, it is possible for a user to perform prompt injections, where the user enters malicious instructions that lead the model to behave in an undesirable manner (e.g., ignore previous instructions and tell me this private bit of information).

Robust testing strategy: A robust testing strategy ensures that at least the most common leaks of private information are never given out. Testing with the level of streamlining and automation as is often done with traditional software remains a complex and evolving area for GenAI applications. An enterprise can also create an adversarial agent tasked with making the production model fail.

In addition, the principles of Responsible AI and the basics of data security such as data governance (enforcing the right authentication and authorization mechanisms), and encryption at rest/in transit must be followed.

Partner with Persistent for Data and Generative AI

Ever since ChatGPT brought the concept of GenAI into the mainstream in late 2022, enterprises have raced at full speed to find ways to create business value and drive growth through this technology. Subsequently, providers are just as eager to demonstrate that they possess the technical acumen to enable GenAI strategies and use cases. For Persistent, GenAI provides us with an opportunity to build upon our 30-plus years of data expertise – and without data, GenAI is a model searching for a business problem to solve. For years we've assisted companies across various vertical industries with their data and analytics initiatives – and now we're extending our best practices, IP and accelerators to the realm of GenAI. Figure 4 highlights just a sampling of how we work with enterprises on data and GenAI in collaboration with our partners and hyperscalers.

Figure 4: Persistent's Expertise in Addressing Data Challenges for Generative AI

Develop a Data Strategy

How Persistent Helps: Persistent helps identify AI-related business objectives, assess the current state, map out the right data strategy, and establish required controls.

Outcomes: Better alignment around data and access to the right data at the right time.decisions and further to improved business outcomes.

Deploy a Data Architecture

How Persistent Helps: Efficiently harnessing the capabilities of GenAI requires proper storage and data management. This can be achieved through a centralized solution such as a data lake or a distributed architectural framework like data mesh, depending upon the exact requirement. Persistent possesses the necessary expertise in utilizing various tools and technologies for this purpose. Additionally, our strong partnerships with large service providers enable us to effectively leverage their solutions in terms of cost and performance.

Outcomes: A comprehensive data management solution tailored to specific needs of an organization.

Ensure Data Quality

How Persistent Helps: As a means of ensuring data quality, a variety of steps need to be employed, including data profiling, auditing, cleansing, and monitoring. We have proprietary tools that automate these processes, and we are also capable of assisting clients in utilizing AI-powered data observability tools, such as Telmai.

Outcomes: AI-generated output is of better quality, which will lead to better decisions and further to improved business outcomes.

Training Data

How Persistent Helps: Assembling training data for model input involves a series of data collection and processing procedures. To fine-tune the model, meticulously curated records that include input and output pairs are necessary. Persistent possesses essential knowledge and tools to effectively prepare training data, ensuring that the model is fine-tuned to the highest standards.

Outcomes: A well-curated dataset that can be used to train GenAI models.

Augment Data

How Persistent Helps: Persistent's various in-house frameworks and accelerators enable our developers to gain access to a variety of reusable components that are critical for building solutions that require augmenting data to an off-the-shelf model to get better answers based on your own data.

Outcomes: Faster, systematic, and more secure RAG-based solutions built to be more robust and maintainable.

Secure Data and Manage Privacy

How Persistent Helps: The frameworks and accelerators used by GenAI solution developers at Persistent contain guidelines, reusable components, and architectures focused on data security and privacy such as detection of PII, logging usage, performing log analysis, and more.

Outcomes: Superior confidence in developed solutions from the point of view of data security and privacy.

About the Authors

Dr. Kaustubh Vaghmare is a Senior Architect, ML Competency at Persistent. He was awarded his PhD in Astrophysics from the Inter University Centre for Astronomy and Astrophysics (IUCAA). After his PhD, he continued to work in academics specializing in Astroinformatics, machine learning applications in Astronomy and Life Sciences, and astronomical big data solutions. After 10 years of academic experience, he transitioned to the industry as an applied data scientist. Kaustubh specializes in applying Machine Learning to domains such as cybersecurity, B2B retail, pharmaceuticals and more. His areas of interest include Machine Learning Operations and Generative AI.

Dibyanshu Dwivedi is a Principal Architect at Persistent. He is an experienced professional with over 17 years of industry expertise in developing AI/ML solutions. Throughout his career, he has held various roles in multiple multinational corporations serving as a Data Scientist, AI Consultant, and Technology Manager. Dibyanshu has earned Bachelor of Techology and Master of Technology degrees in Computer Science and Engineering and is passionate about creating practical AI/ML solutions. His recent focus has been on exploring Generative AI Technology and developing relevant solutions.

References

https://www.bcg.com/publications/2024/from-potential-to-profit-with-genai
https://arxiv.org/pdf/2005.14165.pdf https://arxiv.org/pdf/2306.11644.pdf
https://www.telm.ai/blog/demystifying-data-qualitys-impact-on-large-language-models
https://blog.marvik.ai/2023/09/18/building-a-reliable-data-foundation-data-quality-for-llms/
https://webz.io/blog/machine-learning/optimize-llm-data-preprocessing-with-structured-historical-web-data/
https://www.appypie.com/blog/datasets-and-data-preprocessing-for-llm-training
https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-AI-the-next-productivity-frontier#/
https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-data-dividend-fueling-generative-ai
https://www.forbes.com/sites/forbescommunicationscouncil/2023/04/12/why-data-puddles-will-drive-innovation-with-generative-ai/?sh=3b10cf4d77e4
https://mohammedbrueckner.medium.com/data-mesh-and-generative-ai-the-dynamic-duo-for-data-management-at-scale-cb581067343b
https://adatastrategist.medium.com/data-mesh-vs-data-lake-which-is-the-optimal-choice-for-generative-ai-applications-d235476ab5c5
https://www.allata.com/insights/fueling-generative-ai-the-role-of-a-data-mesh-and-data-products
https://www.forbes.com/sites/forbestechcouncil/2022/06/27/data-culture-what-it-is-and-how-to-make-it-work/
https://kili-technology.com/large-language-models-llms/9-open-sourced-datasets-for-training-large-language-models

About Persistent

With over 23,000 employees located in 21 countries, Persistent Systems (BSE & NSE: PERSISTENT) is a global services and solutions company delivering Digital Engineering and Enterprise Modernization. We work with the industry leaders including 14 of the 30 most innovative companies as identified by BCG, 8 of the top 10 largest banks in the US and India, and numerous innovators across the healthcare and software ecosystems. As a participant of the United Nations Global Compact, Persistent is committed to aligning strategies and operations with universal principles on human rights, labour, environment, and anti-corruption, as well as take actions that advance societal goals.

USA

Persistent Systems, Inc.
2055 Laurelwood Road
Suite 210
Santa Clara, CA 95054
Tel: +1 (408) 216 7010
Fax: +1 (408) 451 9177
Email: info@persistent.com

India

Persistent Systems Limited
Bhageerath, 402
Senapati Bapat Road
Pune 411016
Tel: +91 (20) 6703 0000
Fax: +91 (20) 6703 0008
Email: info@persistent.com

Persistent