Case Study

brillio

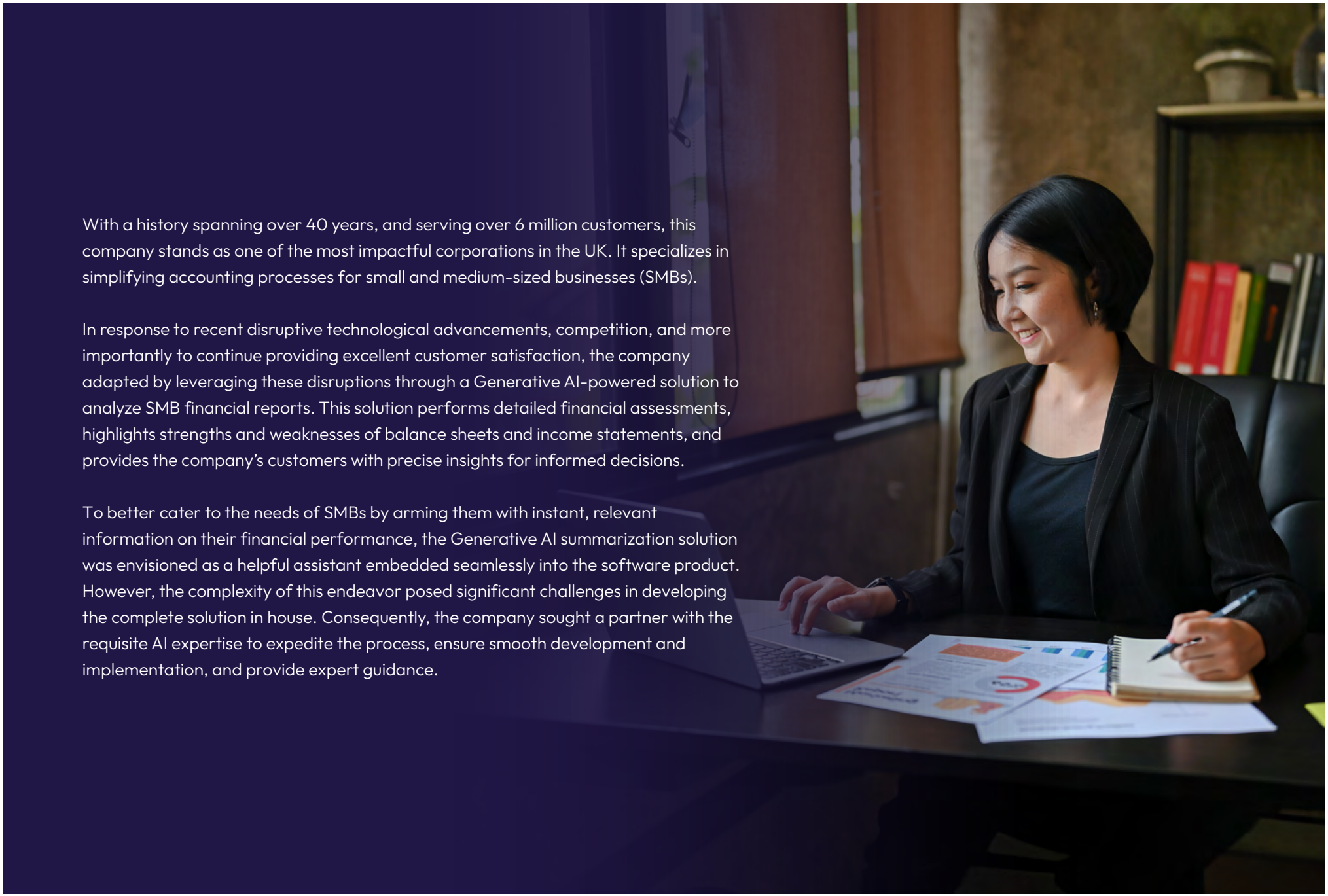
GenAI-powered SMB financial reporting insights

Leveraging Gen AI to transform financial analysis and decision-making for a leading ERP software corporation

With a history spanning over 40 years, and serving over 6 million customers, this company stands as one of the most impactful corporations in the UK. It specializes in simplifying accounting processes for small and medium-sized businesses (SMBs).

In response to recent disruptive technological advancements, competition, and more importantly to continue providing excellent customer satisfaction, the company adapted by leveraging these disruptions through a Generative AI-powered solution to analyze SMB financial reports. This solution performs detailed financial assessments, highlights strengths and weaknesses of balance sheets and income statements, and provides the company's customers with precise insights for informed decisions.

To better cater to the needs of SMBs by arming them with instant, relevant information on their financial performance, the Generative AI summarization solution was envisioned as a helpful assistant embedded seamlessly into the software product. However, the complexity of this endeavor posed significant challenges in developing the complete solution in house. Consequently, the company sought a partner with the requisite AI expertise to expedite the process, ensure smooth development and implementation, and provide expert guidance.



Harnessing Azure Open AI and crafting a roadmap from discovery to deployment

Having successfully delivered similar engagements for clients across industries, including Gen AI intelligent enablement and automation chatbots, GPT-enabled applications, interactive Azure Open AI bots, and LLM Gateway platforms for democratizing Generative AI usage, and featuring the expertise needed to tackle the level of complexity required for this project, Brillio emerged as the ideal partner.

Additionally, Brillio's suite of in-house accelerators enabled expedited implementation, driving increased productivity while the scalable architecture fostered reusability across use cases, enhancing UI/UX, and increasing performance.

Following comprehensive workshops in the discovery phase, Brillio compiled a list of requirements and conducted data assessments and profiling to identify relevant attributes and key ratios from financial statements that needed to be included.

The LLM model evaluated the financial data based on six financial ratios selected as indicators of key business operations. It then responded with identified strengths, weaknesses, and overall conclusions to facilitate faster decision-making.

The following architecture outlines the sequential steps to use the Azure LLM Model within a web app created with FastAPI and Langchain libraries. The web app includes an API endpoint that allows users to interact with the LLM model.

User Request and Authentication: The Gen AI solution uses machine-to-machine authentication that generates bearer tokens valid for 8 hours. Users submit the request after selecting the financial data on the software, which gets transformed to a JSON payload that is sent to the web app and thereafter to the LLM models.

> The solution was designed using Azure Open AI, processing financial data encapsulated in JSON payloads and sending queries to the LangChain LLM Model via API calls.

JSON Payload Preprocessing: Error handling and data validation compare the structure of the received JSON payload against the expected format. If the input JSON deviates from the predefined structure, the API triggers an error message specific to the mismatched elements. The preprocessing module identifies whether the payload is a balance sheet or income statement, extracting language and legislative details for further use and mapping the payload based on legislation.

LLM Model Execution: Based on the identified JSON payload type, the LLM is triggered for financial analysis. For balance sheets, it extracts strengths and weaknesses based on metrics like solvency ratio, debt ratio, and liquidity ratio. For income statements, it analyzes metrics such as economic profitability, return on equity, and revenue generation status. A dual model approach employs a randomized load optimization strategy, dynamically directing requests between two similar-sized LLM models for optimal workload distribution.

Language Processing and Response Verification: For payloads in French, German, and Spanish, the invoked LLM generates a response in English, which is then converted to the preferred language. A response structure check module ensures the generated response conforms to the desired structure.

Logging: All components, such as original JSON payloads, generated response prompts, significant JSON keys, and hash values for logging/caching, are stored in Azure Cosmos DB for archival and retrieval. Upon subsequent submissions of identical JSON payloads, the system retrieves the cached response related to the hash value from Cosmos DB.

LLMOps: The Main branch serves as the primary codebase. Upon each change or feature addition, a new feature branch is created from Main, and all development work occurs locally in Visual Studio Code (VS Code) within these feature branches. After development, a Pull Request (PR) is initiated from the feature branch to the Main. Two levels of approval are required for code stability and functionality. Once the PR receives the necessary approvals, it is merged into the Main branch, triggering subsequent deployment processes. The merge event triggers a GitHub Action for automated deployment to the designated Azure Web App.

Infrastructure as Code (IAC): Terraform, an infrastructure as code tool, is used for safe and predictive provisioning and easy management of cloud resources. Terraform is employed for provisioning new resources or configuring changes (and checking for any setting changes in already deployed resources.

Terraform deployment

The integration of the Terraform script into deployment pipelines ensured consistency, resource testing, and automated provisioning. In GitHub Actions, the process began with the installation of the Terraform CLI, followed by its initialization using the main.tf file (Terraform Init).

After initialization, a validation check was performed on the main.tf file to ensure the correctness of the code, facilitating seamless provisioning (Terraform Validate). Following the validation, a blueprint was generated according to Terraform rules for provisioning new resources or modifying existing ones (Terraform Plan).

Lastly, once the plan was in place, Terraform initiated the process as outlined in the generated plan (Terraform Apply). This integration streamlined the deployment process, enhancing reliability and efficiency.

Simplifying Financial Metrics for Informed Decision-Making

The deployment of the Gen AI solution empowered business owners to effortlessly interpret essential financial ratios such as Solvency, Debt, Liquidity, Average Supplier Payment Period, Profitability, and Return on Equity. This simplified the understanding of critical financial metrics, enabling more informed decision-making. Small and medium-sized businesses (SMBs) are enabled to prioritize operational efficiency over complex financial intricacies, thanks to Generative AI reports that streamline the decision-making process.

The system's monitoring capabilities ensured careful examination of the responses generated by the language model (LLM). This oversight empowers the continual improvement of the quality and effectiveness of the LLM's output, aligning it with the business objectives. The analysis and report generation features served as a roadmap for optimizing the model's performance, ensuring it met the evolving user needs and expectations.

The caching mechanism enhanced system efficiency by generating a unique SHA-256 hash value for each payload from the LLM and storing it in the database. Upon subsequent requests with identical payloads, the database is queried using the hash value, ensuring the reusability of previously generated responses. This process not only reduced operational costs but also saved bandwidth by retrieving analyses from the database more quickly than awaiting a fresh response from the LLM.

The rate limit feature curbed the influx of organizational payloads, ensuring smooth operation. In future stages, personalized quotas based on usage can be tailored to further optimize costs. This transformation reflects a pay-as-you-go model, where resource usage harmonizes with personalized needs, leading to wise and economically sensible use of resources. Additionally, this feature helps to prevent DDoS attacks, enhancing the system's security.

Finally, for load balancing, a dual-model approach ensured incoming requests were seamlessly cascaded between two concurrent models. This dynamic allocation optimized load distribution, with each successive payload transitioning smoothly to the fallback model. The dual models worked together to improve performance and handle requests with efficiency, creating a responsive system. Moreover, this seamless transition not only optimized current loads but also enhanced scalability, allowing the system to expand to meet growing demands.

Reduced operational costs

Enhanced scalability

Optimized resource utilization

Improved performance and efficiency

> The solution provided comprehensive financial analysis from the perspective of a CFO, enabling business leaders to grasp core strengths and weaknesses efficiently.

ABOUT BRILLIO

Brillio is one of the fastest growing digital technology service providers and the partner of choice for many Fortune 1000 companies seeking to turn disruption into a competitive advantage through innovative digital adoption. We help clients harness the transformative potential of the four superpowers of technology: cloud computing, Internet of Things (IoT), artificial intelligence (AI) and mobility. Born digital in 2014, we apply our expertise in customer experience solutions, data analytics and AI, digital infrastructure and security, and platform and product engineering to help clients quickly innovate for growth, create digital products, build service platforms, and drive smarter, data-driven performance. With 17 locations across the U.S., the UK, Romania, Canada, Mexico, and India, our growing global workforce of nearly 6,000 Brillians blends the latest technology and design thinking with digital fluency to solve complex business problems and drive competitive differentiation for our clients. Brillio has been certified by Great Place to Work since 2021.

https://www.brillio.com/
Contact Us: info@brillio.com