A raft of risks is pointing customers toward on-premises generative AI

Marcus Schneider, Deputy Head of Global Portfolio Management, Fujitsu

On-premises generative AI is suddenly big, says Marcus Schneider, Deputy Head of Global Portfolio Management, at Fujitsu. Organizations want the astonishing upsides of today's AIs but are concerned that public cloud models may come with too many risks attached.

Do you remember the astonished realization by business leaders — after the launch of ChatGPT — that generative AI overturns most assumptions about the future of their organizations? If you can think of the right question, then there is a pretty good chance AI will give you an answer that takes you somewhere new and exciting. You can even ask what the right question might be.

What also became clear not long afterward was that relying on public Large Language Model (LLM) AIs, like ChatGPT, Bing, Claude, and Bard, on public cloud platforms comes with a raft of risks. While experiences in the public cloud around feature richness and velocity of innovation have been outstanding, organizations are now — quite rightly — worrying about accuracy, bias, intellectual property leakage, data privacy, emerging regulation and compliance, unknowable legal risks, and spiraling costs, all of which will limit their use of the public cloud.

That's quite a long list of concerns, adding up to a fundamental trust issue for AI.

As a result, many businesses are exploring or already implementing on-premises generative AI. Indeed, according to a recent report, the choice of where customers plan to run generative AI is split almost exactly down the middle in terms of public cloud versus on-premises and edge.

We can confirm the trend – Fujitsu is already working with customers and our partner ecosystem on this. As a result, we already have co-created an on-prem generative AI solution. This solution is also available on our AI Test drive, on which we do have several POCs running together with customers from companies of different sizes in various verticals across Europe.

The root of the problem

Large Language Models rely on trained on massive but finite training data sets. That opens up risks about potential accuracy, such as putting generalization over specificity, lack of verification, and no source of truth. Perpetuating or even amplifying biases is another worry. Any biases in the training data can cause the model to generate biased content.

Beyond the risks of using public generative AIs, public cloud platforms can pose security risks such as data privacy breaches, IP leakage, and operational risks.

On-premises is more precise

Bringing a model back inside an organization's perimeter means that companies can be much more specific about arrangements for security, privacy, data sovereignty, and DR, for example. They can also be much more hands-on about what is actually in the model, reducing the risks of bias and false or out-of-date information.

Organizations can also customize them for specific use cases. They can ensure that the AI is trained on high-quality, diverse, and representative data. And continually update it with new data to improve its accuracy, corroborating the outputs of generative AI with other trusted sources, and educating users about the strengths and limitations of the AI.

Data sources can be much more targeted and include pools that would not be available to a public AI—for example, information held in Teams or on an intranet or extranet. Data about products and services can be withdrawn or updated as they change over time. They can even create models that are ring-fenced to the needs of specific business units or departments. For example, legal teams will want to get advice from an AI where regulation gets primacy over marketing considerations.

In practical terms, this puts users in a position to ask questions like: "Based on the most current versions of our sales presentation and marketing strategy, what are likely to be the most persuasive messages to deliver to the top 10 customers in the fastest growing verticals in Q2?"

All in all, managing generative AI in-house provides flexibility and oversight that cloud-based options may lack.

A more sustainable approach

Ultimately, all these options must be acceptable to customers and society. We've discussed trust regarding accuracy, bias, privacy, and security. But environmental impact will likely become just as prominent in the coming months. The exponential growth in parameters that power AI's impressive capabilities comes at an energy and cooling cost that may prove unsustainable.

However, most enterprise use cases do not require massive models. On-premises hosting allows for tailoring model size, performance, and cost to an organization's needs. Thoughtfully customized on-premises solutions can be right-sized rather than taking a monolithic, one-size-fits-all cloud approach. This aligns operational efficiency with broader ESG commitments.

As concerns mount about generative AI's energy and Cooling needs, on-premises solutions offer a more sustainable way forward.

Getting started is getting easier

With all these upsides, it's little wonder that on-premise generative AI is attracting so much attention.

Getting on board is becoming much easier. Fujitsu, for example, provides reference architectures to optimize infrastructure configuration and sizing, and a complete stack together with our partners, depending on the use case.

Do you have an idea for a private enterprise specific use case for an LLM where you need advice on how to proceed? Maybe you want to finally understand the content and implications of the hundred or so contracts your operations team has open right now? Obviously, you don't want to upload those documents into a public environment. We provide comprehensive advice on what works and what doesn't and how best to implement your idea.

Fujitsu's Test Drives of the DX Innovation Platform provide state-of-the-art infrastructure and consultancy support to help understanding complex requirements, as well as validate and evaluate the data to build business cases and select the right infrastructure. Additionally, Fujitsu's private GPT solution can be tested as well.

To request a contact or to register for a Test Drive go here: https://mkt-europe.global.fujitsu.com/DX-Innovation-Content-Hub.

Marcus Schneider
Deputy Head of Global Portfolio Management, Fujitsu

Marcus has been working for Fujitsu in various functions, including running the Data Protection Business and managing the Storage Development team.

Before becoming the Deputy Head of the Global Portfolio Management, he was responsible for the European Portfolio.