

The Edge LLM AI

WITH THE REVENUE EXTRACTION GATEWAY

When ChatGPT was released, it was nothing short of a revolution. It swiftly captured hearts and minds around the globe, becoming an integral part of our digital lives. Its ability to understand and generate human-like text was unparalleled, making it a beloved tool for many. But beyond its immediate appeal, ChatGPT sparked an explosion in Large Language Model (LLM) research, applications, and products. Researchers around the world were inspired to push the boundaries of what was possible with LLMs. New applications emerged, ranging from customer service to content creation, and a plethora of products were developed, each harnessing the power of LLMs in unique ways. This white paper delves into the exciting world of edge LLMs, a field that owes much to the pioneering work of cloud-based predecessors like ChatGPT.

ChatGPT

You

Can you give me a 4-day itinerary to experience the wildlife in the Australian Outback

ChatGPT

Absolutely! Here's a suggested 4-day itinerary to experience wildlife in the Australian Outback:

Day 1: Alice Springs

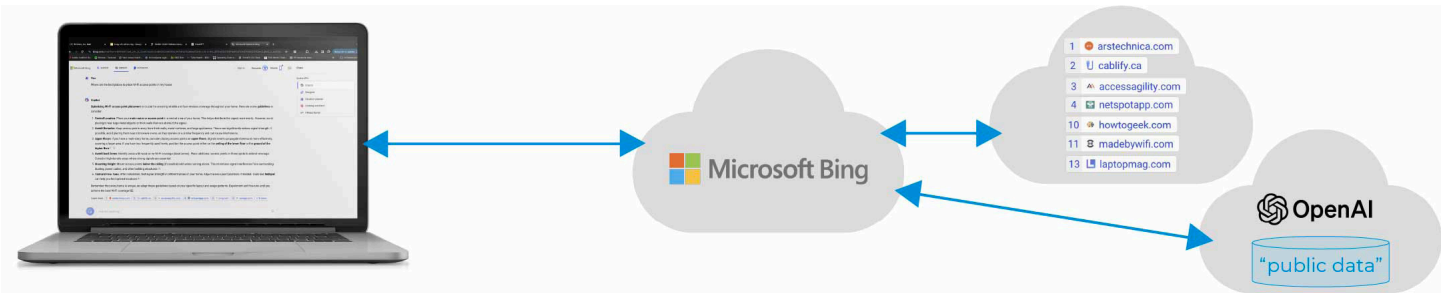
- Morning: Arrive in Alice Springs, the gateway to the Outback. Visit the Alice Springs Desert Park to see native wildlife like kangaroos, wallabies, and emus.
- Afternoon: Explore the Alice Springs Reptile Centre to see a variety of reptiles, including snakes, lizards, and crocodiles.
- Evening: Take a guided nocturnal tour to see unique Outback creatures like bilbies, echidnas, and night birds.

Consumer cloud-based LLMs, such as ChatGPT, are primarily designed to work with public datasets. Their strength lies in their ability to answer a wide array of questions, such as “Create an itinerary for an adventurous vacation in Australia for animal lovers.” These models excel at such tasks because the information used to generate the answers is entirely public and changes infrequently. This means they can leverage vast amounts of data to provide detailed, accurate, and creative responses. The use of public datasets also ensures that the responses are not only relevant but also, when used in conjunction with Retrieval Augmented Generation (RAG), up-to-date, making cloud-based LLMs a powerful tool for information retrieval and content generation.

When LLMs are applied to business contexts, they encounter a completely different set of requirements with respect to the dataset. Business datasets have three key characteristics: confidentiality, size, and rapid changes. Confidentiality is paramount as businesses often deal with sensitive information that must be protected. The size of the dataset is another crucial factor. Businesses generate vast amounts of data daily, and LLMs must be capable of handling this volume efficiently. Lastly, the rapid changes in business data present a unique challenge. Unlike public datasets that change infrequently, business data can change rapidly, reflecting the dynamic nature of business operations. Therefore, business-oriented LLMs must be designed to adapt quickly to these changes, ensuring they provide accurate and timely insights.

| | | |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| Data Privacy | Will my data be used to train LLMs that will be accessible to my competitors in the future? | How will I replicate my data segmentation and role-based access controls? |
| Data Size | How large is my proprietary dataset? | How rapidly does my dataset change? Would it be acceptable to generate answers given out of date information? |
| Recurring Cost | How much will I pay per month for my data at rest in a simple storage service? How much will decryption cost every time I access it? | How much will I pay each time I hit the LLM? Have I accounted for the context that I will need to feed the LLM in order to get reasonable results? |

In the intricate tapestry of business operations, proprietary datasets are the golden threads that weave together a company’s unique value proposition. The guardians of these treasures are data custodians who are typically Certified Information Systems Security Professionals (CISSPs). The mere thought of these datasets being transported to an off-site storage, potentially uncontrolled and vulnerable, is akin to opening Pandora’s box. The risks are not just about theft; they extend to the misuse of data, including the alarming possibility of them being used to train the next generation of LLMs. It’s a chilling prospect that underscores the critical importance of robust data governance and stringent security measures in our increasingly data-driven business landscape.



| Provider | Model | Context | Input /1k Tokens | Output /1k Tokens | Per Call | Total |
|--------------------------|---------------|---------|---------------------|----------------------|-----------------|---------------|
| Chat / Completion Models | | | | | | |
| OpenAI / Azure | GPT-3.5 Turbo | 16K | \$0.0005 | \$0.0015 | \$0.0011 | \$0.11 |
| OpenAI / Azure | GPT-4 Turbo | 128K | \$0.01 | \$0.03 | \$0.0213 | \$2.13 |
| OpenAI / Azure | GPT-4 | 8K | \$0.03 | \$0.06 | \$0.0440 | \$4.40 |

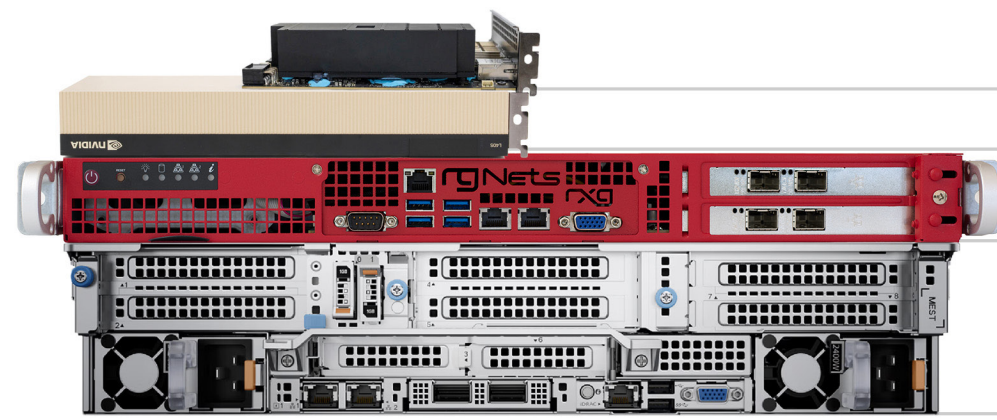
** The table does not include the cost of storing the proprietary dataset, decryption the data, nor the multiple round trips that will almost certainly be required to get something useful out of the system, plus the potential requirement for side chats.*

The size of the dataset is also a significant concern when it comes to business applications of LLMs. Large datasets, when sent to cloud storage, can incur substantial costs. For instance, storing 20TB of data at rest in a simple cloud storage service can cost over \$2,000 USD per month. But the expenses don't stop there. If the data stored in the cloud is encrypted for security purposes, businesses also need to account for the CPU consumption used for decryption, adding another layer of cost. Furthermore, any changes to the data can incur transport fees, which can quickly add up given the dynamic nature of business data. Therefore, the size of the dataset not only poses technical challenges but also financial ones, making it a critical factor to consider in the deployment of LLMs in business settings.

In the ever-changing landscape of business, data is like a river, constantly flowing and evolving. This presents a unique challenge for LLMs. Unlike static scenarios, businesses thrive on the pulse of the latest information, be it sales data, medical records, or IT telemetry. Cloud-based LLMs face a conundrum - they either need a direct lifeline, an API access to the proprietary data, which could be overwhelmed by the data deluge, or they need to mirror the data, constantly replicating it, a task akin to capturing a river in motion. Both scenarios underscore the intricate dance of real-time data access and the complexities it brings to the fore in the application of LLMs in business.



On-Premises LLM Solution

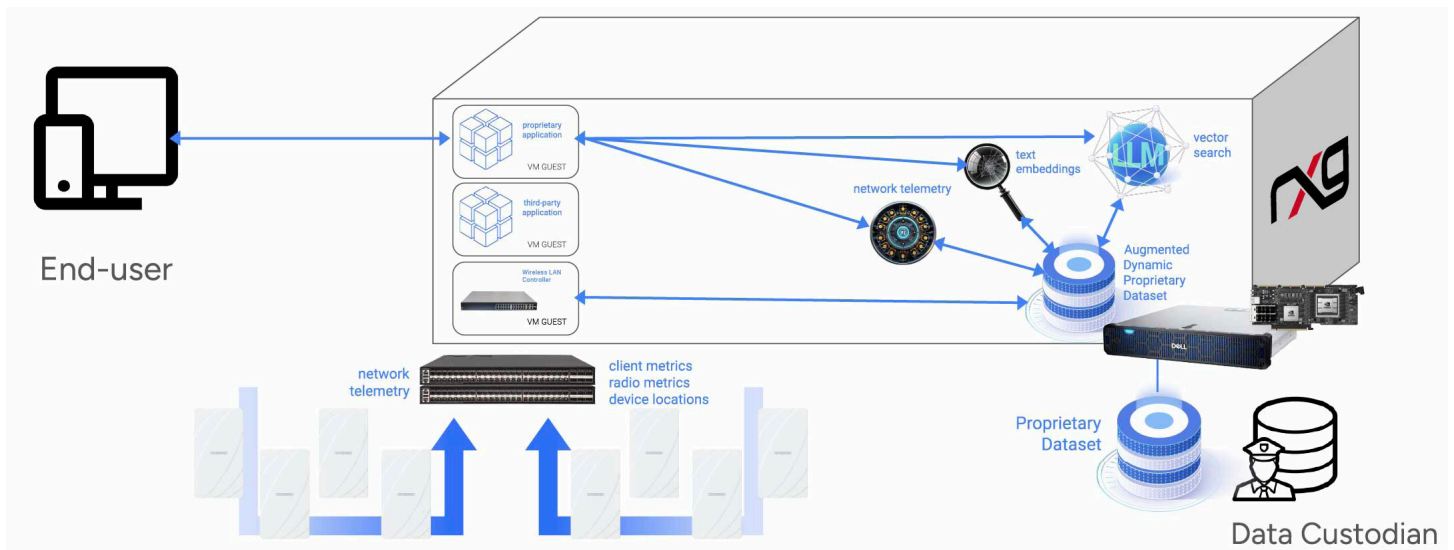


NVIDIA BF3 DPU

NVIDIA L40S GPU

RG Nets rXg

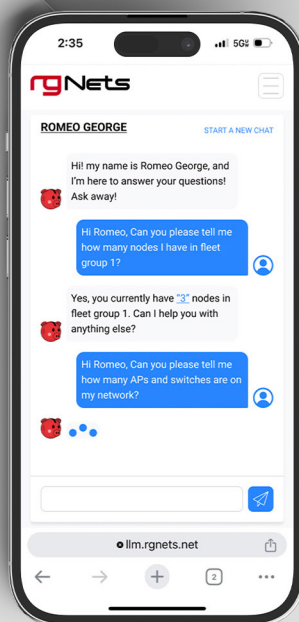
Dell XR7620



Our Edge LLM solution elegantly addresses the three major challenges of confidentiality, data size, and frequent changes. Operating on the edge, or within a private cloud, the LLM has direct access to as much storage as the business is willing to provide. This storage is owned and controlled by the business, safeguarded by in-house information security professionals. It's entirely feasible to link the LLM to a read-only copy of the live storage of the proprietary dataset. This ensures that the LLM always has access to the most recent, up-to-date data, effectively turning the challenges into strengths. This is where its power lies - it brings the intelligence to the data, rather than bringing the data to the intelligence.

This solution is a symphony of the latest advances in RAG technology. It brings together a suite of cutting-edge techniques, including, but not limited to, vector databases, document preprocessing, side chat strategies, and multifaceted prompt engineering into a flexible, user-friendly, and cost-effective package. Running on local GPUs pulling from local and remote datasets at the edge not only addresses the fundamental challenges of data confidentiality, size, and rapid changes, but also offers a significant cost advantage. Unlike cloud-based solutions that bill per generated token, with our Edge LLM solution, the cost is fixed. You invest in the inferencing hardware once, and the only ongoing cost is the power and connectivity. This makes our offering not just technologically advanced, but also economically efficient, truly bringing the power of LLMs to the edge.

The Edge. Redefined.





www.rgnets.com
sales@rgnets.com
316 CALIFORNIA AVE
RENO, NV 89509

