www.pipelinepub.com
Volume 21, Issue 4

Really Smart GenAI Phones: Ecosystem Risks

By: Mark Cummings, Ph.D.

GenAI started running on publicly accessible data centers full of GPU's, mostly Nvidia chips. Now we are on the cusp of LLMs (Large Language Models — the Software engines of GenAI) moving to edge devices, Notebooks, and Smart Phones. This technical transition. Concerns about IP leakage, plus cost factors, will make this a disruptive change, indicating a need for caution in making large financial commitments.

Economic Risks

In November of 2024 Bloomberg, in an article titled "Wall Street's elites are piling into a massive AI gamble,“ reported on a dinner meeting organized by senior people at Morgan Stanley with leaders of the largest private equity funds. Bloomberg estimates that it will require trillions of dollars to build upcoming GenAI data centers, associated nuclear power plants, and communications networks. The Morgan Stanley organizer reportedly told the group that the financing required was beyond the capacity of the banks and suggested partnering with the large private equity funds.

It appears that this economic analysis depends on two key assumptions: 1.) inference will continue to be run on large data centers accessed via telecommunication networks; and 2.) there will be a large number of teams simultaneously training new fundamental models. Both assumptions seem open to question.

Inference at the Edge

In the form of GenAI prevalent today, there are two very different types of processing: Inference and Training. Inference is the process of using a fully trained LLM. It is called inference because the trained LLM is asked a question. In answering the question, the LLM "infers" the answer based on the question, the block of "attention" data that provides context, and its training. In most publicly available LLMs the inference response time is in the one second range with thousands of inferences performed per minute. Although more processing intensive than a database look up, inference is many orders of magnitude less processing intensive than training. Once trained, some GenAI systems can be run on 2023/24 generation notebook computers.

By Q4 2025, Nvidia is expected to have a chip designed for PC's. In the same time frame, Apple is expected to have an M5 chip fully optimized to run high performing LLMs on its notebooks. By late 2026 it is reasonable to expect similar chips capable of running LLMs on Smart Phones. The appearance of these chips will bring into question the first assumption of continuing to run inference primarily in large data centers.

Declining Demand for Training

It appears that the number of groups creating new fundamental models is currently declining. A recent analysis of IaaS (Infrastructure as a Service) costs found that the cost of renting one Nvidia GPU for an hour has been declining. At the peak it was $8.00/hr. It has declined to less than $2.00/hr which is $0.50 below cost.

The largest portion of this cost decline is a result of the reduction in the number of teams creating new fundamental LLMs. The analysis recognizes that a small portion of the price decline is due to relief of a temporary Nvidia chip shortage. As parameter set sizes have grown to the high tens of billions of parameters, training costs have become prohibitively high. Many teams funded to create new fundamental models have switched to either tuning existing LLMs or creating medium sized LLMs. The analysis indicates that the number of teams globally creating fundamental LLMs is less than 50 and may be continuing to decline.

Without high and growing training demand for data center resources to do training, the demand for GenAI data center capacity may shrink. This brings the second assumption into question.

How Training Effects Demand for Large Data Centers

An over simplified view of training is that it is a series of inferences. Training of a fundamental model starts with a model that is constructed with a fixed architecture and randomly assigned parameter values (currently in the 80 to 100 Billion parameters per LLM, expected to grow into the Trillions). Then, the LLM is asked a question to which the answer is already known (in the training data). The answer the LLM provides is compared with the known answer and the difference between them is measured. This difference is often called a cost function.

Next the model is run backwards to find the part of the model that if changed will produce the most change in the answer. Once that parameter is found, it is manipulated, the LLM is run again, and the cost is measured. This is repeated until a minimum cost is achieved. This process is iterated for each of the many billions of items in the training data set. Then, iterated with combining costs from different parameters runs.

Thus, training is extremely processor intensive. Even with the most powerful GPUs it can take many weeks of 24/7 processing using an entire large data center of the most powerful GPUs to train a billion parameter LLM. With just one step down in the power of the GPUs, the many weeks can become many months. This is the reason that generations of a particular fundamental model are typically 6 months apart.

Fundamental models can be further trained. This further training is done by adding new material to the training data set and conducting more sets of iterations. This further training is sometimes called tuning. Typically, the new training data is several orders of magnitude smaller than the fundamental training data set. Therefore, tuning is less processor intensive. Tuning may be done to strengthen the LLMs capabilities in general, or it may focus on one particular area or domain.

Data provided in inference questions or in the attention blocks can leak out to other users. To avoid this IP leakage, some large companies and government institutions are considering creating private GenAI systems. Companies are concerned with proprietary data leakage. Governments are concerned with other types of information leakage. It also appears that some are considering creating their own custom fundamental LLMs. That is, doing their own training of models from scratch. If this comes to pass, it would reverse the decline in the number of active teams doing fundamental training. Such a reversal could create sufficient demand for the projected trillion dollar investment in GenAI data centers.

Another alternative is for each of these organizations to take an existing fundamental LLM and fine tune it by adding training with their proprietary data. Such tuning would be adding a few percentage points of new training data. Thus, if this approach is followed, it would not result in the demand for large numbers of data centers.

It appears highly unlikely that, with the exception of a very few government agencies (intelligence? military?), organizations would have the staff talent necessary for, and could justify the expense of, creating their own fundamental LLM. Thus, most who try would fail.

Likely Outcome

Pending the next disruptive innovation of the order of the one that triggered GenAI to start with, the technology/market forces described above indicate the following.

The overwhelming majority of inference accesses will be performed on Edge resources. Starting with notebooks first appearing in late 2025. Then, adding local LAN accessed servers. Followed shortly thereafter by inference locally on Smart Phones in late 2026.

For security reasons, organizations may restrict access to LLMs tuned with their proprietary data. The restriction will be in layers. The first layer will be access control (password, multi factor authentication, etc.) to local servers and data center implementations. For more critical data, physical access controls may be deployed. That is, access only available from physically secure end points. For the third and highest level of security, the LLMs may be air gapped. That is, have no internet or similar network access and be physically isolated along with their access points.

Other than for security reasons, Inference on large organization private data center resources will be primarily driven by applications that require very high levels of performance. In the public GenAI data center space, demand will be similarly driven by the need for very high performance.

There will be promises of IP leakage prevention. But organizations are likely to be skeptical of these claims. So, public usage will generally be used only for applications where privacy and security are not significant considerations.

These high performance LLMs will continue to grow in size and complexity. The costs involved in coding, training, testing, regulation, liability, will also continue to grow. The result will be a continued reduction in the number of teams developing new foundation models. Thus, a reduction in the number of large data centers required to perform training.

An entire supporting ecosystem will be developed in the process of evolving to this Edge dominated use of LLMs.

Economic Danger

The danger is that if the ecosystem developed doesn't correctly anticipate how things will actually play out, there can be stranded assets.

If trillions of dollars are invested in new GenAI data centers, a significant portion of the resulting assets could become stranded. That is, there would not be enough demand to generate the returns to pay off the loans used to create them. Such a situation would cause pain for those directly involved. Pain for society in general if a significant number of data centers are involved and the financial failures ripple through the general economy.

It takes time to build a new data center and an associated nuclear reactor for power. First comes the design and permitting process. Then, constructing the data center building. Both of these can take up to a year each. Filling the data center and connecting it to network resources can be done more quickly. Building a nuclear reactor takes longer than building a data center. Initially, many of these new data centers may try to run on grid-supplied power. The risk can be somewhat limited by financing done in tranches tied to each stage. But if much of this infrastructure is built and the users don't show up, there will not be enough revenue to pay back the loans.

A possible scenario is the construction of 200 $5 billion data center/network/nuclear reactor complexes. If one such complex targeted on public inference and training services were started in January of 2024, permitting and design completed in January 2025, building/network construction completed in January 2026, limited service starting in June 2026. Full service starting in 2029 with the completion of the nuclear reactor.

Meanwhile, the Nvidia PC and Apple M4 chips come out in Q4 2025. By June 2026, inference on Edge systems has growing market share. Demand for data center inference starts to plateau and then drop. While a drop in demand for public training services is triggered by the shrinking number of teams developing new fundamental LLMs, plus security concerns, there might be time to cut the financing for the nuclear reactor. But a large percentage of the assets financed would be stranded.

Some will argue that these assets can be targeted at other sources of demand. Others argue that the overall demand for inference will grow so fast and so high that even with percentage reductions, the overall demand will be sufficient to make these investments economic. Those arguments may be correct, but maybe not. There is also the threat of another disruptive technology breakthrough. Thus, the need for caution.

Conclusion

Advances in chips designed to run GenAI on Edge devices like notebooks and smart phones herald a move of inference off data centers. Technology and market forces are reducing the number of teams developing new fundamental LLM=s and the concomitant demand for data center resources to support LLM training. This transition will be disruptive in a variety of ways. This likely scenario indicates a need for caution in making large financial commitments.