LLM and electricity

David Leitch

2025-01-01



image-20250109092107532

## Warning Will Robinson

This is my first look at this topic and it’s only a desktop/literature review over a limited holiday time period and has not benefitted from management discussions. It could be absolutely wrong. Also the cost metrics and replacement cost metrics represent the work of a beginner. Analysis skill is no substitute for actual knowledge. And the latter will take time. If nothing else the note has a few interesting photos.

## Summary

I guess the data centre market breaks down between centres used to train and run LLMs and the rest of the market. Much of the excitement around data centres comes from associated growth of LLMs. However being int he LLM space requires very deep pockets and a willingness to keep upping the ante. Traditional data centres may have more modest growth but the business is more stable and predictable with lower depreciation and quite possibly higher returns on capital. Still much of this note focusses on data centres for LLMs.

There are few barriers to entry in the LLM model space, demand is likely to grow strongly but unevenly. Model capabilities continue to expand and LLM models are capable of completing more complex tasks which further expands the use cases. A second stimulus to demand is the 1,000 % reduction in processing cost per token over the past couple of years.

Still unlike most computer activities the ongoing operating costs in terms of energy consumption and tie up of computer time from LLMs is very significant.

Just as there are few barriers to LLMs so there are few technical barriers to the creation of data centres whether of hyper scale (let’s say power consumption greater than 100 MW) or smaller. The real barriers are likely to be cost of capital and quantity of capital to deploy. In the end for the platform businesses this is likely to mean the hyper scale data centre business in the Western world will end up in the hands of a few major providers with a long tail of medium and small players. Equally as ever management expertise is going to be important. What to invest in, when to do it will differentiate the winners from the also rans.

Returns on capital will depend on the nature of the revenue contracts, for data centres devoted to reasoning models scale can be a differentiator but there is also the need to keep up the ongoing chip advances. It seems to me that data centre could have very fast technical obsolescence just now. That’s because hardware technology is changing rapidly and most of the capital cost of a datacenter devoted to LLMs is in the hardware as opposed to balance of plant like power supply, backup power supply, cooling system etc.

That LLM hardware gets out of date within a year or two just now. Out of date means new chips that can do more computations per unit of energy become available. In this note I somewhat grandiosely introduce the metric of Data Centre power efficiency which accounts for not only the PUE but also the processing power of the hardware expressed relative to an NVIDIA H100. Not Data centre energy consumption per token processed.

Equally the design of components that support the chips, and enable communication between them also evolves. Likely too is that at the top of the heap clusters of hyper scaled data centres will be established enabling some task sharing.

In the case of Australia it appears that the major players either have constructed or have actually under construction about 1.8 GW of data centre capacity including some centres that would qualify for the low end of hyperscale. Based on capacity that’s already starting to approach the present largest industry in Australia for energy demand, aluminium smelting. What you can’t see from the public data is the capacity utilisation and therefore the underlying returns. As best I can work out from desktop analysis the equipment in the Australian data centres is not of the power of modern H100 or higher spec centres and this may account for valuations only 1/3 of the cost of an H100 or higher spec centre.

There remains a strong debate about the ultimate returns on capital and productivity boost from AI. Amara’s law holds that technology breakthroughs are over hyped in the short term and under estimated in the long term. Equally the big platform companies all identify that the risks of not having a competitive product are much larger than the risks of over investing.

So for the time being expect to see the total data centre capacity rise, hyper scale centres going from 100 MW perhaps as high as 500 GW of power demand per site maybe even more and I expect the industry, like so many, to consolidate around an oligopoly of sorts but also with a long tail.

For Australia the issues are:

* We have a high number of startups willing to rent data centre capacity, but does it have to be in Australia?
* The firmed energy requirements of a maxed out data centre are rigorous as are supporting infrastructure requirements for data transmission and liquid cooling. Just having competitive bulk VRE is not going to be enough.
* Can we capture a share of the South East Asian market for LLM? For instance if you were an Indian company and didn’t like your electricity supply in India would you consider Australia as a place to train and operate your model?

## Energy use in Data Centres

A forecast by semianalysis (who build their analysis from the bottom up) as of March 2024 shows an increase in electricity demand of about 900 TWh between 2024 and 2030 representing around 3% increase relative to 2023 global demand of about 23,000 TWh. This chart is adapted from the original and recreated using Claude AI to my satisfaction.

|  |
| --- |
| Figure 1: Data centre global power consumption. Source:“semianalysis.com” |

There are many factors that go into this but personally I am partial to the blue line which shows 1,500 TWh by 2030 compare with NEM consumption of say 250 TWh by that year or China’s present demand of maybe 6,500 TWh in total. Below I look through lots of data, noting the economics aren’t that strong; the use cases are not that well defined, and the power efficiencies of data centres are increasing quickly reducing energy requirements per square metre. If I have read the information and done the sums correctly and assuming that the current standard AI GPU is something called an H100 then the new Blackwell series is more than twice as efficient as measured by output per Kw even though the required input power is also higher.

|  |
| --- |
| Figure 2: Relative chip efficiency.Source:ITK based on semianalysis |

What i think this means is that the required power won’t grow as quickly as the work done by AI and this will contribute to overall lower cost per token. So the output can double but input power doesn’t change. Still output is expected to quadruple and more so clearly more power is going to be required. Globally demand is around 300 TWh in 2024 or about 34 GW of demand.

Equally according to seminanalysis there is a 3X cost improvement in going from H100 to B300. Given that prices to users are already down as much as 1000% over the past two years, it’s going to be interesting to see where we end up.

### A data centre energy efficiency metric

Power unit effectiveness (PUE = Total Facility Energy/IT Equipment energy) is only part of the story. We also need a metric for processor effectiveness which I notionally think of as tokens processed/KWh. I therefore propose:-

* Data centre power efficiency = PUE \* TE where;
  + PUE = Total Facility Energy/IT Equipment energy and;
  + TE = Tokens/KWh (expressed relative to say an H100 baseline)

The reason for including the token efficiency is because the economics in terms of energy, the main operating cost, have to be measured against the amount of work done which in turn is measured by the number of tokens, input, reasoning and output processed. Or so I think.

## Data centre segmentation and the implications for Australia

Publicly available data with just a little bit of estimation for Equinix suggests that there is already close to 2 GW of data centre capacity in Australia. At a 100% capacity utilisation factor that suggests 16 TWh of consumption or 7% of annual supply. But capacity utilisation is likely much less.

|  |
| --- |
| Figure 3: Data centres Australia. Source: company websites |

But maybe not all of it will be used near term.. I could easily imagine capacity utilisation at little over 65%. However there isn’t much evidence for that. To the extent that the capacity in Australia is used for more conventional data purposes as opposed to AI then then revenue streams, particularly for CDC where Govt is a major customer, may be much steadier.

As I say I am a neophyte when it comes to this stuff but if you build a hyper scale data centre full of H100 chips and the market has moved to B300 then the economics may not look all that solid. And that seems to be the risk in the AI field. One of many risks.

Just because AI is increasing demand for power doesn’t mean that there will be a data centre explosion in Australia. Globally SemiAnalysis segments the market:

* Retail data centres a few MW located in cities.
* Wholesale - 10-30 MW. Value proposition to customers is scaleability over time, often built in stages;
* Hyperscale: Self built by hyperscalers for exclusive use, individual buildings 40-100 MW and typically multiple connected buildings eg a Google 300 MW centre

### Australia replacement cost metric

A 100,000 H100 system has a total IT equipment cost of about US$4 bn according to semi analysis and a power consumption of maybe 150 MW. This much is more or less reliably estimated.

|  |
| --- |
| Figure 4: 100k H100 datacentre IT cost. Source:Semianalysi |

In the semianalysis models that humble individual like me can see, the assumption is that the IT is “co located” and the colocation cost is stated as about 110/ $US/kW/mth. Colocation means that the IT equipment is installed at the site of some one that supplies data centre infrastructure, that is the building, the cooling, the power supply, the step down transformers.

If I work the Semianalysis cost metric through to building a Australian data centre replacement cost metric off very limited foundations, I get to around A$62 m/MW which is far in excess of the Airtrunk sale price and NextDC’s trading enterprise value. Nevertheless here is the worked table.

|  |
| --- |
| Figure 5: Bottom up replacement cost. Source:ITK based on Semianalysis |

By contrast available valuation metrics for Airtrunk and NextDC are less than 1/3 of that replacement cost metric.

|  |
| --- |
| Figure 6: Australian data centre market values. Source: Yahoo Finance and media |

I do not have a good explanation of the difference and without having the benefit of due diligence discussions with datacentre management my theories are probably wrong. However my first guess is that the equipment in most Australian data centres is not of the same complexity or power as Nvidia H100 and therefore costs a lot less and valuations reflect that. Equally the number of MW varies as more are being built.

## The cost of using LLMs has declined dramatically

The past twelve months have seen a dramatic collapse in the cost of running a prompt through the top tier hosted LLMs.

In December 2023 (here’s the [Internet Archive for the OpenAI pricing page](https://web.archive.org/web/20231214190432/https://openai.com/pricing)) OpenAI were charging $30/million input tokens for GPT-4, $10/mTok for the then-new GPT-4 Turbo and $1/mTok for GPT-3.5 Turbo.

Today $30/mTok gets you OpenAI’s most expensive model, o1. GPT-4o is $2.50 (12x cheaper than GPT-4) and GPT-4o mini is $0.15/mTok—nearly 7x cheaper than GPT-3.5 and *massively* more capable.

Other model providers charge even less. Anthropic’s Claude 3 Haiku (from March, but still their cheapest model) is $0.25/mTok. Google’s Gemini 1.5 Flash is $0.075/mTok and their Gemini 1.5 Flash 8B is $0.0375/mTok—that’s 27x cheaper than GPT-3.5 Turbo last year.

I’ve been tracking these pricing changes under my [llm-pricing tag](https://simonwillison.net/tags/llm-pricing/).

These price drops are driven by two factors: increased competition and increased efficiency. The efficiency thing is *really* important for everyone who is concerned about the environmental impact of LLMs. These price drops tie directly to how much energy is being used for running prompts.

A different source offers the following figure:

|  |
| --- |
| Figure 7: Cost of 1 million tokens. Source: Foundation Capital (Ashu Garg) |

The incredible fall in cost increases the use cases as does the increased input token length. My own personal experience from using AI is that already it’s become indispensable and it may well lead to the productivity increase that the world needs. And on other comparisons the cost has dropped by 1000% in a bit over 2 years.

## Roy Amara’s law

Apparently, and in my case, we tend to overestimate the effect of a technology in the short run and under estimate it in the long run. Equally the consensus of all the hyperscalers, that is Microsoft, Meta, Apple, Google is that the risks of under investment are much less than the risk of over investment. And that goes to explain the hyperscaler arms war. if you don’t have the capacity you aren’t in the game.

Next I turn to an excellent mid year piece from Goldman Sachs economist Joseph Briggs. [Briggs](https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gen-ai--too-much-spend,-too-little-benefit-/TOM_AI%202.0_ForRedaction.pdf) estimates that generative AI will boot US productivity by 9% and US GDP by 6% just in the next decade and result in 300 million full time jobs that will likely be automated within a decade.

Briggs showed that even in the USA as of early 2024 the share of business using generative AI was still low.

|  |
| --- |
| Figure 8: AI adoption. Source:Goldman Sachs |

And also showed that 60% of today’s jobs did not exist in 1940. In fact I’m surprised its only 60%.

|  |
| --- |
| Figure 9: employment by age of industry. Source:Goldman Sachs |

## The state of the art according to Simon Willison

One of the things that distinguishes between the echelons of the finance industry is working out who to listen to on a particular topic. I don’t know which echelon I belong in but I do think Simon Willison is good on LLMs and through him one finds other references.

Let’s start with a table of significant model releases during the year. This table itself was partly built using code developed by DeepSeek v3.

|  |
| --- |
| Figure 10: 2024 LLM releases. Source:Simon Willison |

The table shows how many model releases there were and that the standard of the models is quite close. .

Next i basically paraphrase Simon Willison’s end of year piece [“Things we learned about LLMs in 2024”](https://simonwillison.net/2024/Dec/31/llms-in-2024/) .

Simon’s first point is that many models are now doing as well as or better than GPT-4. He references the industry leaderboard where industry people try different models on different problems and report the results.

|  |
| --- |
| Figure 11: LLM performance. Source:Simonon Willison |

Like Simon, for programming I personally use Claude 3.5(3.6) Sonnet and find it very useful. Indispensable in fact, because with it I undertake and complete programming tasks well beyond what I would try without it. Eg getting an SMS whenever the 5 minute spot electricity price goes over $1000 .

But the more general point is that lots of models are up there and the scores are close at the top.

The next point is that progress on the models is broad as well as deep. Specifically:

* Models now accept much longer inputs (token lengths in the jargon). Last year 8,192 tokens was the standard. Now it’s 200,000 and Gemini(from Google) accepts 2 million tokens. Longer inputs allow a lot more user problems to be addressed
* Models now accept live video and audio input. Eg you can point your phone camera at your pantry and ask CHAT GPT through speaking to the phone to come up with some recipes.
* Inference reasoning models are on the rise. That is the model re evaluates each step it takes to solve a problem and can get better results. The model uses more inference rather than more training.
* Speaking of training, arguably one of the biggest advances was in the model developers learning how to use synthetic data to train models rather than further scraping the finite heap of increasingly proprietary protected web data.

## LLM training limits and overcoming them

Reading the literature, some doubt emerged in 2024 about the ability of models to keep improving. To date models have followed scaling principles. The more data they are trained on the better they perform. However clearly that process has limits in that there is a limit to the amount of training data. It grows a bit each year as fresh content is created but the fresh content doesn’t grow anywhere near enough to sustain optimistic views of model improvement.

So basically as far as I can understand in 2025 at least LLMs will improve in two main ways:

1. Models will use **inference** a lot more. My interpretation of this is that the model will explain each step it takes and that explanation is more or less input into the model for the next step. This is computationally more expensive but is already producing results. Equally there can be a hierarchy of agents within a broader model each seeded with a different prompt.
2. And I think this may well turn out to be the biggest source of improvement, but it’s a sheer guess, models will be trained on “**synthetic data**”. Again from my reading examples of synthetic data might be translating an english language data set into say Japanese and then training the model on the Japanese set or similarly if there isn’t enough HTML data to train a model, you can get one model to translate Python to an HTML equivalent and then train the new model on the HTML as well as the Python. As the LLMs themselves become more sophisticated so their ability to create synthetic data grows. Obviously there are challenges like building in errors. Synthetic data can also be used to overcome existing diversity datasets. In healthcare synthetic data sets can be used to train models without impinging on confidentiality. There is already an industry for the supply of synthetic data eg. Gretel.ai

Equally beyond straight LLM training the other obvious trend is that models can accept input from a wider variety of sources. All the major providers now allow smart phone camera video as an input source and phone audio. ie you can talk to the LLM with your phone and with the video on showing for instance a list of ingredients in the kitchen and ask your ai provider to suggest a recipe. Then you can ask it to improve the recipe and see where you end up. In my opinion this wider variety of inputs will tend to improve the models, but this is just my opinion, in the same way that travel tends to broaden the mind of us humans. Watch this [Simon Willson youtube](https://www.youtube.com/watch?v=mpgWH9KulUU) to see the brand new Gemini 2.0 model in action. Well worth it.

## Energy use, investment and build out

There are a couple of preconceptions I have that ahead of writing this:

* There is going to be a very capital intensive wave of data centre investment;
* The returns on that investment will be low;

As always with new industries and new tools finding the appropriate revenue model is important. It’s no good just having a shiny new toy it has to be monetised. Eg lasers.

Let’s look at what’s going on but in contrast to starting from a top down perspective, I’m going to reference some bottom up stuff done from what seems like another great [industry expert](https://semianalysis.com/2024/10/14/datacenter-anatomy-part-1-electrical/).

* Retail data centres a few MW located in cities.
* Wholesale - 10-30 MW. Value proposition to customers is scaleability over time, often built in stages;
* Hyperscale: Self built by hyperscalers for exclusive use, individual buildings 40-100 MW and typically multiple connected buildings eg a Google 300 MW centre

|  |
| --- |
| Figure 12: existinbg hyper scale data centre. Source:semianalysis |

|  |
| --- |
| Figure 13: Nvidia H100 power. Source:semianalysis.com |

A hyper scale datacentre requires all the electrical equipment to step down incoming voltage from say 230 kV down to that required by servers. This will typically be done in steps, down to 33 kv input to a building and then further stepped down.

And if someone was building a 100K H100 cluster (which I don’t suppose anyone will build now the the new B200 and B300 are becoming available). That electricity cost is based off 150 MW of power. As Semianalysis notes that 150 MWhas to go into a single building, similar to a purpose built aluminium smelter. This means that you need HV transmission access, a switch yard and then all the power conversion equipment to step the voltage down from over 100 Kv to the voltage required by the pods. Even in Australia there are not so many places where you just hook into 150 MW of input power. Spreading out the 100K cluster quickly gives away compute advantages due to the wait time for data transmission between buildings.

|  |
| --- |
| Figure 14: 100k cluster annual power cost. Source:Semianalysis |

If you want to read up on the gory details start with [semianalysis note](https://semianalysis.com/2024/06/17/100000-h100-clusters-power-network/).

Note that the above figure is already out of date, as new designs, particularly new communications between superpose are emerging all the time.

Backup generators currently diesel but potentially batteries are required at every LV transformer. Look at these backups!

|  |
| --- |
| Figure 15: Backup generator. Source:semianalysis.com |

### 500 MW per facility may not be enough

The latest facilities push some big numbers. The Google 300 MW centre apparently can reject close to 200 MW of heat! Incredible number:

|  |
| --- |
| Figure 16: Cooling and backup. Source:semianalysis.com |

The efficiency is measured by PUE (Power Usage Effectiveness) and is as low as 1.1 for the above facility. Without liquid cooling capacity per square metre falls by about 35%. In higher temperature conditions liquid cooling requires about 1 L/kWh.

Google has been building their data centres close to each other allowing some interconnection.

So now Microsoft and OpenAI aim to out Google Google by building very large interconnected campuses.

|  |
| --- |
| Figure 17: Connections. Source:semianalysis.com |

# Appendix. Approach to analysis, a generalist has strengths a specialist may not

This note is at best an early sketch of what’s going on in data centres and the progress of LLMs (Large language models). Although I am a “bush” python programmer and spend a lot of time both programming and using AI to assist with the task I would not claim any more than a generalist knowledge. So this is my first attempt at looking at this industry both as a segment of electricity demand, as a data centre market, and to glance through the economics of LLM. Also this is just a desktop study and has not had the benefit of a review by anyone let alone discussion with centre management. As such the entire note could easily have entirely the wrong end of the stick, so to speak.

As a generalist analyst though one’s training both enables and encourages analysis of industries as they develop. I trained as an analyst and accountant. In my first job out of academia I was the only analyst at a small stockbroker in the 1980s. I covered every stock. One day BHP the next Pacific Dunlop.

Over time some industries became the specialities, building materials, banks for a while, infrastructure and utilities. My perspective is that industries have similarities and differences and the broad exposure equips me to look at new industries. I might be wrong.

A defining theme of modern industry is technology disruption. Not every industry has it, building materials barely changed in 30 years, but deregulation in utilities, innovation in finance (stapled securities) and genuine technology moves (solar, batteries, evs) result in a constant need to re evaluate seemingly settled opinions about the value of particular businesses.

When the pace of technology development is high and the industry is new the range of opinion about what will happen is wide. And that’s because potentially the outcomes are themselves open. Or seem so. In fact they probably aren’t. Certain general rules will likely apply. What are these rules?

* If the business is very profitable it will face competition;
* If the market looks like it will be large new competitors will enter; James Hardie found a magic sweet spot, a business, fibre cement, that could support a lot of growth in the USA for one company or maybe two, but not so large as to warrant the attention of really big players. AI is much, much larger and has attracted already many serious, deep pocketed competitors;
* As the market grows, costs tend to come down and often develops, particularly in tech, into a platform style winner takes all model. In any event a few large players often end up dominating the industry. Often the first innovator doesn’t have the skills to stay the course. Sometimes though, think Tesla, they do. What is for sure is that the skills required to be a long term profitable player are much different to those required to be the early disruptor. For instance in stock markets in new industries there is a cynical suggestion that it is often not until a business gets its third owner that a return on capital is earned because the price has adjusted, the scales fallen from the eyes. See eg Victorian power stations post privatisation.