WHITE PAPER

Buying AI for the Enterprise

By Phil Anderson

National Sales Manager: Digital Business Solutions at Datacentrix

SERIOUS ABOUT PERFORMANCE

Passionate about value

Table of Contents

1. Buying AI for the enterprise – cutting through the complexity

In a very short space of time, the world has progressed from a single breakthrough AI tool (ChatGPT) to a veritable tsunami of offerings from every single vendor in the market. The focus has shifted from simple personal productivity pitches to enterprise-defining business use cases, which impact every facet of a company's operations. The evolution of what is available to purchase and deploy has been significant, and pricing models have evolved at a rapid rate – with no end in sight.

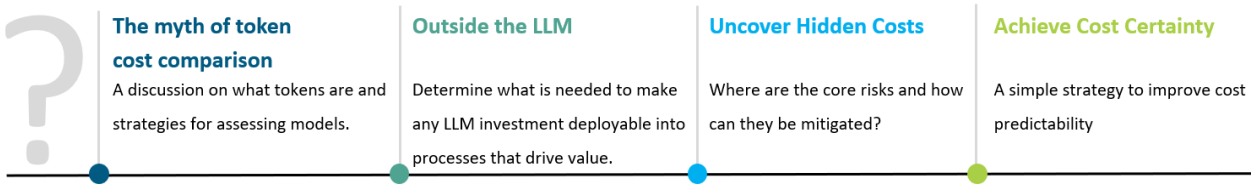
This causes several significant challenges for buyers, raising numerous questions. How can different solutions be compared? What is needed to make this work? Where are the hidden costs? How can certainty of operational costs be achieved? And just to be very clear, the software vendors are not making this easy. In fact, quite the opposite. It is not in any software vendor's interest for the comparison to be straightforward, or this swiftly becomes a race to the bottom as seen with other enterprise software offerings over time.

For the purposes of this white paper, the focus will be on where the most value is available to enterprises and that is not in the personal productivity tools space. These are purchased on relatively simple, user-based structures, so the decision is more straightforward and clearer, with familiar licensing models as used for other types of software categories.

So, why are productivity tools the area of least value? It's simple. In most of these cases it is extremely difficult to directly determine whether making a person more productive adds to business outcomes. For example, if an automotive manufacturer saves a group of its employees an hour a day on emails, will it really help them to sell more cars? How is this value tracked and proven?

The focus of this paper will be on enterprise use cases, integrating directly into business process, which require the use of a large language model (LLM). This is where the highest return can be seen in the market and the real business impact can be ascertained via standard return on investment (ROI) and time to value (TTV) calculations with which every organisation is familiar.

This will be broken down question by question in the following sections.



2. The myth of token cost comparison

A 'token' can briefly be defined as the unit that model providers use to measure consumption, and therefore attribute charges to a business's use of their model. For example, when a query is sent to a model, it will consume tokens to input the query and consume more tokens to generate the output. The same applies to training a model; this will also consume tokens. The token has been adopted by the AI industry as a standard unit of measure.

Many buyers have started with comparing the cost per token of the different LLMs available. However, this is a 'rabbit hole' to be avoided, and here's why.

Put simply, token cost is not directly comparable across models. Not even nearly. Different models use tokens differently. While the **term** is standard, the **application of tokens to use cases differs drastically** across vendors. Here are some examples when it comes to model inferencing:

- In some models, a token will allow the LLM to consume 10 characters of text, in others 15, or perhaps eight.
- Certain models consume tokens for spaces, some do not.
- Some models cache the first letter of every word, while others cache full words, and some do not cache any at all.

These factors significantly impact the number of tokens consumed for a use case, so the listed rands and cents cost per token becomes largely irrelevant. And this is before addressing model-tuning costs, where some include this and some do not within token definitions, bringing yet another layer of complexity.

So which approach should be taken in LLM pricing comparison? There are three possible strategies:

1. Analyse all the differences in what a token provides against each model under evaluation and create a financial model to assess. Frankly, this is a non-starter. Not only would this approach trap teams in 'Excel hell', but the answer it will generate will likely already be out of date. It will always be behind the curve given the pace of innovation and development present in the market today.

2. Run a specific use case against **all** the models under evaluation and assess actual cost, not token consumption. This is, on paper, a strong approach. However, it is a luxury few companies can afford and may need to be completed for every use case. This is neither practical nor sustainable in practice for all except the largest enterprises.

3. Make peace with the fact that the cheapest model will not always be the right one and focus on the function needed for the specific use case and the fundamentals defined for the organisation in its AI governance and policy. Note: if there is no such policy in place, this is something that Datacentrix can assist in compiling.

Below is a more detailed explanation pertaining to strategy three, as it is the one that Datacentrix favours and recommends for most organisations.

Different models have different strengths and are designed for different purposes. It is important that the right category of model is used to apply to a use case. Some models are generalists, but they are less likely to be as efficient for specific tasks and mostly do not allow any fine tuning. It may be that a generalist model – with little to no tuning – will be the best option, but if the requirement specifically focuses on improving a business through targeted deployments into process, it is unlikely.

So… which capabilities are really needed to serve the use case**?** Summarisation? Classification? Extraction? Question answering? Text generation? Translation? Inference? Conversation management? Sentiment analysis? Entity recognition? Intent recognition? Code generation? Code explanation? Planning and task sequencing? Math and computation? Knowledge retrieval? Multi-modal understanding? Image recognition? Image generation? The list gets longer each day with plenty of capabilities to choose from!

Identifying which capabilities are required to support a business process use case should be the first step in identifying feasible models**,** then the model choice can be narrowed down significantly to reach an answer. The market for different models is constantly growing, but the hyperscalers and OEM platforms have a suite of their own models and allow a 'bring your own model' approach and access to great model libraries, like Hugging Face. Hugging Face is the leading marketplace for open source models and offers a lot of value to this process. It is important to note that model size should also be assessed – smaller models are more efficient and consume less effort to tune.

The reference to AI governance and policy is important as it comes down to **trust** in two aspects. Firstly, in today's deployment models, there are generally not copious graphics processing units (GPUs) available in the cupboard, businesses rely on breaking out to access LLMs via application programming interfaces (API's) into the provider's cloud. It's critical here to be sure that said provider can be trusted with data, that they will keep it safe and not use it to feed other people's insights, perhaps competitors.

Secondly, it is important to be certain that other people's data is not inadvertently being consumed. Where is the data for the training sourced? Is the vendor transparent about their training approach or is there a risk of downstream legal issues? Data sovereignty, regulatory exposure and potential supply chain risk based on shifts in AI export policy are all significant factors in these decisions. **Trust** should **be a** major buying factor for enterprises today.

In summary:

- Token cost is a highly misleading comparator as it is not a consistent unit of measure. Thus, time should not be wasted here.
- Matching use cases to model capabilities and size is essential for maximum efficiency. Avoid using sledgehammers to crack nuts.
- Work out who can be trusted with data and determine whether they match predefined ethical standards

3. What is be needed to make this work?

Currently the obsession in the market is with LLMs, but LLMs only represent the engine of the car being assembled. Engines can be switched out, when necessary, but the chassis and steering need to be consistent and enterprise scale. Essentially, in AI, the surrounding architecture of the LLM is as important, if not more important than the LLM itself.

Why? The model on its own will take input and provide output. How this insight is integrated and orchestrated into the business process is what drives the value. So, what are the prime components to be considered?

1. Integration – in most use cases, there will be a requirement to interact with systems that are already in use within the enterprise, transporting data to and from models. This must be managed through an integration bus using common API standards. If there is no platform in place, this must be considered in any proposed solutions. However, those with existing integration platforms and associated standards already in place should design according to those.

2. Orchestration – this is a significant consideration. How will all the data pipelines, execution of various models, workflows and data sources be managed in association with the wider objective? This is at the core of the 'agentic' movement currently under wide discussion, which combines AI and different automation disciplines, such as robotic process automation (RPA), to co-ordinate agent activity across the enterprise in a single control plane. If an organisation is planning AI deployment at scale over time, this is an absolutely foundational component.

3. Data governance and management – this has been said many times by many people about AI, because it matters. If there is no **proven** understanding of data being provided to the models, then dark times lie ahead. The architecture and method required here cannot be underestimated and need to be made a key focus!

4. Model management – what is the home for all these AI exploits? Using an LLM as standalone for an enterprise is possible but it brings risk, especially in cases where tuning is required. What is the AI management platform that will be used to manage the models and data sources, manage model drift over time and make tuning a more straightforward task? A home is required for this innovation.

In summary:

- The LLM is the engine, it is up to the business to build the car.
- Orchestration is the key to deployment at scale.
- Data quality matters more than ever before.

4. Are there hidden costs?

This is a question asked of every software vendor, and the answer is always the same. “No Mr. Customer, our costs are completely transparent!” And in many cases, they are telling the truth. The problem is that the customer is not really asking about the software costs. What they are really asking is, **“What could go wrong and how much could it cost?”**

In comparison to innovation waves and hype cycles of the past, AI seems to require the highest amount of ‘school fees’ to be paid. While this could arguably be deemed as appropriate due to its high potential return, it is also the one where there have been many more failures than successes, and failed POCs have been much more common than successful ones. It is clear when exploring the numbers that more money is being made across the industry in consulting than software sales right now… and this is adding to the complexity of delivering value in this space today.

The industry is learning, and what should an organisation’s role in that learning be? Learning is critically important, but it is also vital to understand what needs to be spent before gaining value.

One of the largest unknowns is the number of tuning cycles that are needed to get to the perfect model for the use case. It is a truly iterative exercise that requires a level of experimentation to reach the right outcomes. The higher the customisation needed, the more iterations required. This is another contributor to the importance of selecting the right model, with the capabilities closest to the use case!

The other factor in tuning is human. Defining what ‘good’ looks like is extremely difficult when dealing with a technology that can deliver so many broad answers. The more clearly a successful response or answer is defined across the majority of inputs to the model, the easier it becomes for engineers to reach the end point. **A vague requirement in this field costs even more** than a vague requirement in more **deterministic technology.**

To reduce these ‘school fees’, partnership is important. In many cases today, vendors are doing free work to develop intellectual property (IP) and take the resulting offerings to market. Organisations can be smart by identifying use cases that would be re-usable and then using IP negotiation to reduce/negate the experimentation cost. However, allowing vendors to experiment on your data without locking up the IP beforehand, means losing a bargaining chip from day one!

In summary:

- AI’s ‘school fees’ can be high.
- Tuning is iterative and can easily run on past the plan.
- Vague requirements will cause more harm here than in other technology projects.
- Use IP terms to create shared value and reduce cost.

5. How to reach certainty of operational costs

Given that models are consumption based, and the consumption costs are constantly changing, how can operational costs be secured when buying solutions?

With the pace of evolution in the technology, one clear strategy would be to buy the value and not to build it. And this is where 'as-a-service' models come in. These models pass the risk on, and, in an early-stage industry, this is not a bad thing. It is 'outcome-based buying' in the extreme. In other words, it focuses on gaining the business value promised, regardless of the technology used and the number of resources needed. As a side point, this allows the reader to fundamentally ignore most of what has been written above! With apologies of course, but this gem could not have been included at the onset.

It is, of course, possible to buy the infrastructure needed to run these workloads, but this is an option only for those with deep pockets. In the "As-a-Service" approach, the token costs and usage-based logic fall away, but there are different concerns here, such as throughput, efficiency, model portability and, most importantly, total cost of ownership (TCO).

Essentially, a solution provider like Datacentrix is here to take all that complexity discussed off the table for the client. The client wants the outcome, Datacentrix takes away the risk of delivering and running it. It is SaaS packaging for a non-SaaS world.

This model doesn't mean that businesses should stop experimenting and learning themselves – they should most definitely be doing that – but it does mean that they can take the key use cases and dramatically reduce cost of delivery, 'school fees' and all the other risks discussed in this paper. Even if this is not necessarily a long-term strategy, it is a great place to start.

As-a-Service means that the right focus can be placed on business value, while the technology delivery is left to someone else as that space matures.

Summary

- Think about buying a service, not commissioning a project.
- This allows the focus to be centred on business value, not technology.
- Pass risk to the partner, but don't stop learning!

6. Conclusion

To recap the discussion, here are five points to remember:

- **Point 1:** Token cost is a misleading 'rabbit hole'.
- **Point 2**: Avoid the use of sledgehammers to crack nuts – get the model fit right.
- **Point 3:** The whole car needs to be built, not just the engine.
- **Point 4**: Focus on minimising 'school fees' and getting to value fast.
- **Point 5**: Consider as-a-service to pass risk to partners while the market (and the business) matures.

As a closing note… no AI was used in creating this paper, or even the summary. The aim was to write it personally, based on real experience, not what a model dictates. For this piece, the author relied solely on the human brain and not AI (for a change).

Interested in further conversations? Contact Phil Anderson on PAnderson@datacentrix.co.za for more information on our innovative capabilities.