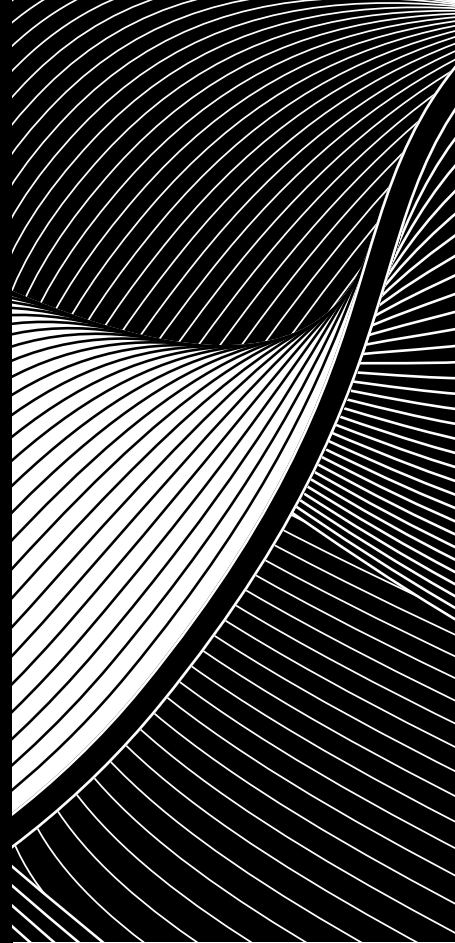


Pathways

From prompts to power: Generative AI and implications for data centre demand

October 2026



3

Introduction

6

Generative AI: What it is and where its electricity goes

10

Workload-based projections: Forecasting demand, not supply

13

Key results: Electricity demand rises ninefold

17

The main levers: Efficiency versus intensity, and the video wild card

23

Implications for investors: The scale of the build and the kind of load it serves

27

Appendix: Generative AI electricity demand – model methodology

Executive summary

Investors in data centres and the power ecosystem supporting them face a difficult question: is capacity being built ahead of AI demand that never arrives? This question is not easy to answer, not least because many of the projections available are based on supply-side signals, such as chip shipments and building pipelines. They are easy to track, but they only describe what the industry is building, rather than what users will need.

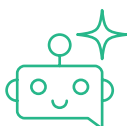
In this paper we have taken a different, demand-based approach, modelling demand from how the technology is actually used. We attempt to address this question from the perspective of generative AI – in our view the subset of the broader AI discussion that is most relevant and attracts most investor time and discussion. Our key findings are as follows.



AI-driven power demand growth is real and significant. Demand is forecast to grow ninefold by 2040, underpinned by generative AI. Generative AI's electricity use rises from about 91 terawatt-hours (TWh) today to roughly 857 TWh in 2040 even with significant hardware and software efficiency gains modelled. In our high case demand reaches 1,219 TWh, thirteen times what it is today.



Investment opportunities abound. Building the data centre and required power infrastructure will require some \$US4.5 to \$US5.6 trillion by 2040, or roughly \$US300 to \$US375 billion a year.



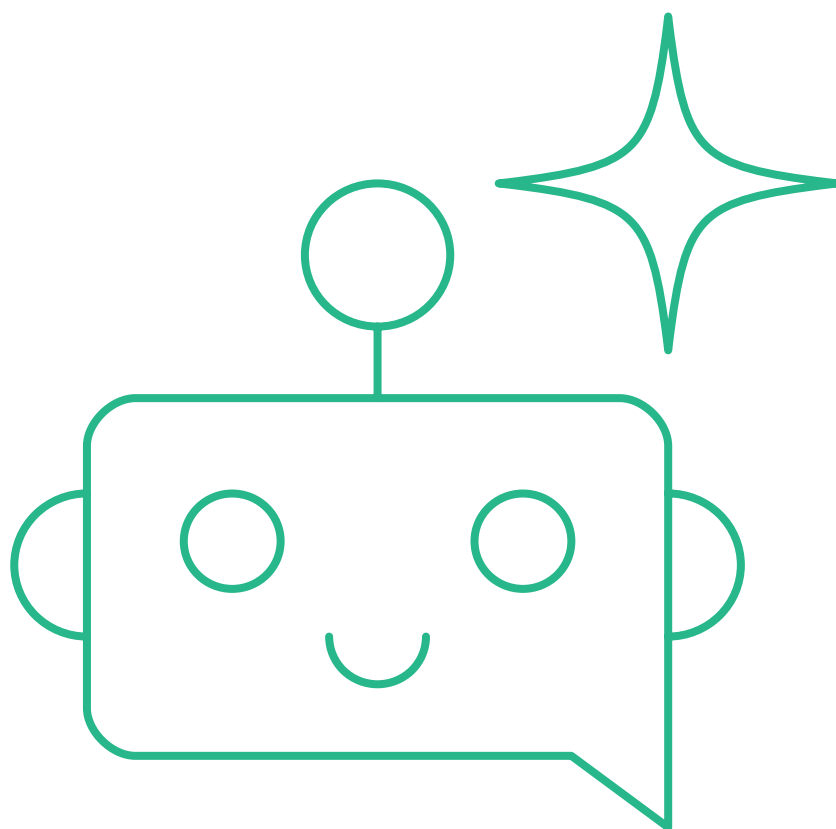
Agents are the future. The proliferation of AI as agents used for tasks currently performed by humans is the critical next evolution. Demand growth is driven far more by intensity of use than breadth of take-up. In our base case heavy users are about a quarter of the total by 2040 yet account for some 70% of the electricity behind inference. Behind that is agentic use: the model works through tasks on its own rather than answering a person, so its output is no longer bounded by what a human can process. As an example, the fast-developing world of humanoid robots – not within the scope of our paper – would add further upside to our power demand forecasts.



A more predictable but less flexible load matters for developers and grid owners. Generative AI at work (“inference”) consumes more electricity over time than generative AI training, and this is a more predictable but less flexible draw on power use. Inference rises from under 60% of demand today to over 85% by 2040 in our base case. It follows the users and usually answers on demand, resulting in use that tracks the working day and week, whereas training tasks can be queued but draw power more erratically and more flexibly.

AI-driven infrastructure demand has been hotly debated over the past few years and continues to capture investors' imagination across the world. Our bottom-up work leads us to conclude that the growth will be significant and real. Permitting and environmental impact may continue to influence the path, and there will be difficult questions ahead on grid availability and behind-the-meter solutions. Nonetheless, large amounts of infrastructure capital will be needed to meet the demand generated by this new technology. The even harder question for infrastructure investors is whether that power can be delivered where and when the load asks for it.

Introduction



Few forces are reshaping the outlook for electricity demand as quickly as artificial intelligence and, within it, one workload in particular: generative AI. It is the technology behind large language models and the tools built on them, and it behaves quite differently from the machine learning that has run in data centres for years.

It is also where the money has gone. Global private AI investment tripled between 2022 and 2025, from \$US113 billion to \$US345 billion, and generative AI accounted for around 70% of that increase.¹ That is already changing load forecasts, grid plans and generation investment across much of the world. Those assets are financed over long horizons, so the question for anyone building or funding them is not just how fast capacity can be built, but how much electricity AI will need over the life of the asset, often the next ten to twenty years.

Today's projections mostly forecast supply. They extrapolate from what can be counted: power contracted, data centres announced, chips shipped. That data is timely and it is real. But they describe what the industry is preparing for, not what will be consumed. Chip shipments are the clearest case. Hyperscalers order well ahead of demand, and there is a backlog, so shipments show what suppliers expect to sell rather than what users will need. That gap is why the risk of oversupply is a live question. Forecasting supply therefore has several shortcomings.

- **Is the demand real?** Supply-side estimates measure commitment rather than consumption, so they cannot tell a build-out that demand will fill from one that has run ahead of it. They cannot, therefore, answer the fundamental question of whether we are in an investment bubble.
- **Which AI?** They cannot separate demand from generative AI that started the current boom from predictive and older machine learning workloads. That matters because the two behave quite differently: the older workloads grow steadily and predictably, while generative AI is the volatile and rapidly growing part the capital is exposed to. Blending them makes the total look steadier than the decision facing an investor.
- **Training or inference demand?** Supply-side estimates cannot distinguish the two, though each places quite different demands on a network and on the assets serving them.

This paper takes a new approach and models demand, building projections from the bottom-up fundamentals of how the technology is actually used: how many people use generative AI and how heavily; how much each request asks of a model; and what serving that request costs in energy. For training, we start from what the leading laboratories spend on compute. Training is a deliberate choice about how much to spend, so that spending is the demand, not a proxy for it.

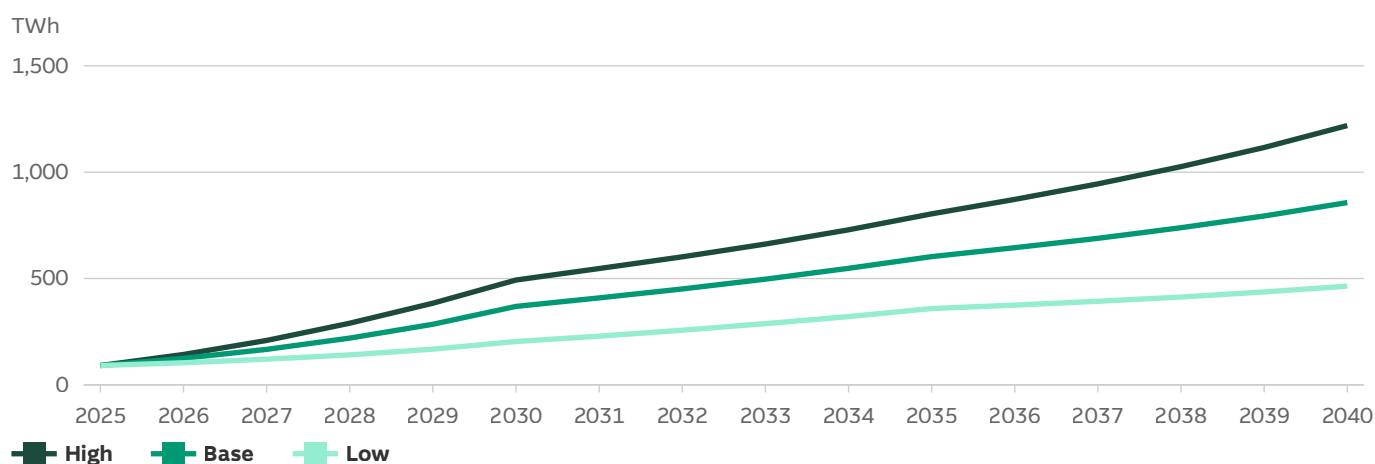
1. Source: Stanford HAI, AI Index Report 2026, April 2026.

Our analysis gives a considerable spread: 857 TWh in 2040 for our base case, within a range of 464 to 1,219 TWh (Figure 1). That range reflects the genuine uncertainty about the technology's long-term trajectory, but our analysis also offers a view of the drivers behind it, which is where the more interesting findings are:

- **Intensity drives the growth, not headcount.** How many people use AI matters far less than what they use it for, with a small group of heavy and increasingly machine-driven users carrying most of the load.
- **The load will likely become more predictable but less flexible.** AI electricity consumption divides between training (building a model) and inference (using it), and the two behave differently. Inference follows the users and usually answers on demand, tracking the working day and week, whereas training can be queued but draws power more erratically. In our base case inference rises from under 60% of generative AI demand today to more than 85% by 2040, so the load becomes more predictable but less flexible as it grows larger.
- **The demand is real and meeting it will require substantial amounts of capital.** Even where adoption stalls and use stays light, demand rises around fivefold by 2040; in the base case ninefold, and at the top of our range thirteenfold. Continuous draw goes from around 10 GW today to between 53 and 139 GW. Adding that much load in fifteen years is not primarily a generation problem – it is a question of where it lands, how quickly it can be connected, and whether the firming and transmission assets arrive with it.

Figure 1.

Key results: Electricity demand in different scenarios



Source: Macquarie Asset Management analysis.

The rest of this paper builds those conclusions in five steps. Section 1 defines generative AI and explains how it draws electricity. Section 2 sets out the approach and the three scenarios. Section 3 presents the results and places them against published forecasts. Section 4 examines what sets the level of demand and explains the key drivers of different outcomes. Section 5 discusses the implications for investors.

SECTION 1

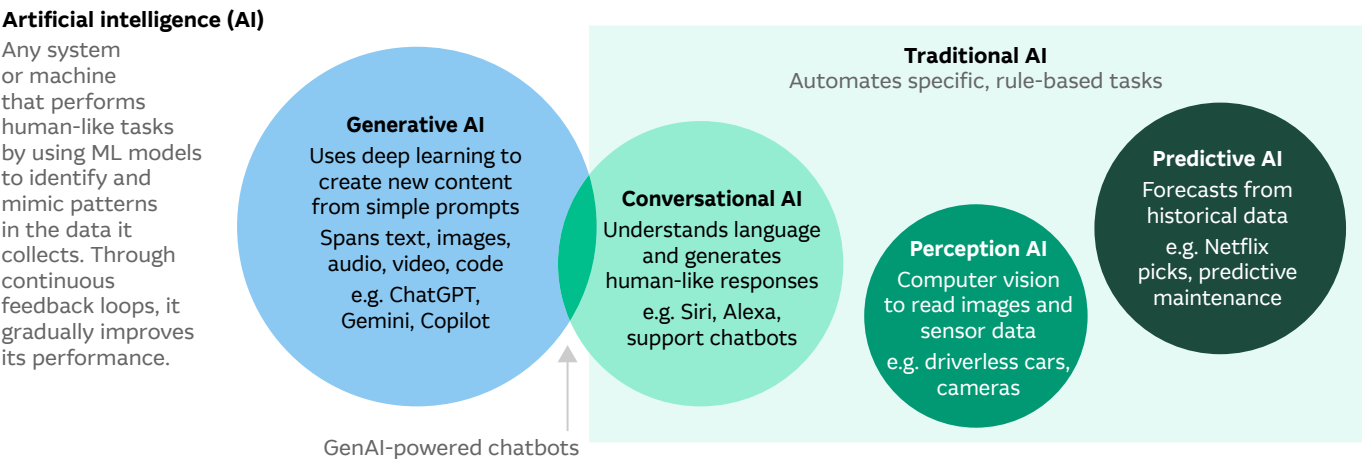
Generative AI: What it is and where its electricity goes



AI is the latest accelerant of a decades-long shift of economic activity into the digital world, and that shift runs on data centres.

Data centres have long done the everyday work of the internet: cloud computing, storage and corporate IT; AI is simply their latest and fastest-growing use. AI in turn is a broad field, spanning recommendation engines, computer vision and forecasting, as well as the generative models behind ChatGPT. Generative AI is one segment of AI, not the whole of it (Figure 2).

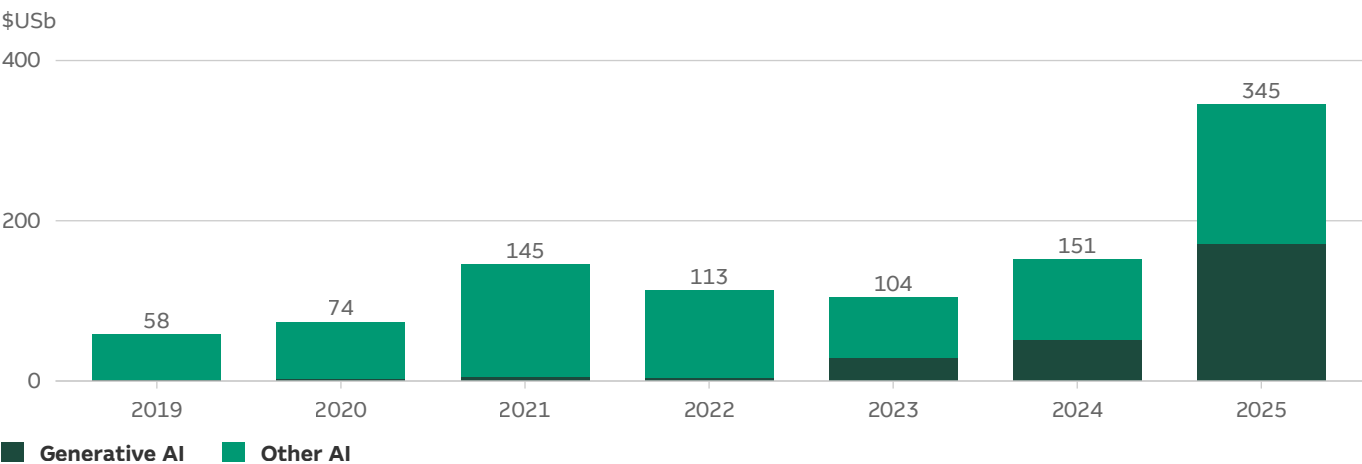
Figure 2.
Generative AI versus other types of AI



Source: Microsoft AI, “Generative AI vs. other AI types”, accessed August 2026; Macquarie Asset Management analysis.

The distinction matters because the data centre figures most often quoted capture one of the broader categories, all data centres or all AI, while the capital is directed at generative AI in particular. Global private AI investment tripled between 2022 and 2025, and generative AI went from under 4% of that to almost half, accounting for around 70% of the increase (Figure 3). It is the part of the picture that is new, that is growing fastest, and that could turn out to have been overbuilt. Much of the debate is nonetheless conducted on numbers that include a great deal else besides. This paper focuses on generative AI specifically and estimates its electricity demand on its own terms.

Figure 3.
Generative AI as share of global private investment in AI



Source: Stanford HAI, AI Index Report 2026, April 2026.

What generative AI is and where its electricity goes

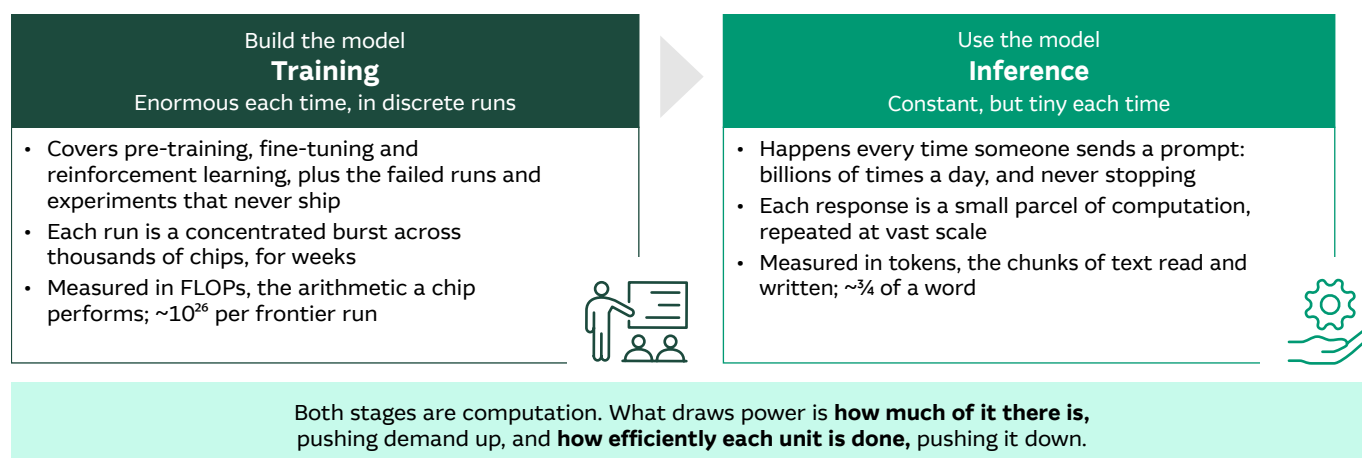
Generative AI is software that learns patterns from vast amounts of data and uses them to produce new content, text, images, audio, video or code, in response to a prompt. The system doing the learning is called a model, and the models behind most of the current interest are large language models (LLMs): ChatGPT, Gemini and Claude are all built on them. Images and video come from a different kind of model, which is why we measure them separately. Providers do not offer a single model but several, ranging from small and fast to large and capable: Claude comes as Haiku, Sonnet, Opus, Fable and Mythos, Gemini as Flash and Pro. A short question is usually handled by a smaller model, while harder work goes to a larger one that costs far more energy to run. Underneath, everything they do is computation running on specialised chips, and that computation is what draws power.

A model's life has two stages, and each runs its own computation (Figure 4):

- **Training – Building a model.** A neural network is exposed to enormous datasets until it absorbs their patterns. For any one model this is a one-off, but across the industry it never stops. In 2025, Epoch AI recorded six frontier models and more than a hundred notable ones.² The real figure is higher still, considering smaller releases fall below the counting threshold as well as failed runs and experiments not reported. Compute in a run is measured in floating point operations (FLOPs). A frontier run now involves around 10^{26} of them; an ordinary laptop running flat out would need something like a million years to get through that many.
- **Inference – Continuous usage.** Every prompt and response is processed in tokens, the small chunks of text a model reads and writes. (A token is roughly three-quarters of a word. Images, audio and video convert into far more.) Each token is a small parcel of computation, individually tiny, but repeated billions of times a day and never stopping. Increasingly, that output is not read by a person at all. Agentic systems, where a model is given tools and a goal and calls other models in a loop, consume their own output as the input to the next step. That removes the ceiling human attention places on response length.

Figure 4.

The two phases of generative AI compute



Source: Macquarie Asset Management analysis.

The two stages behave very differently. Training is a fixed cost: a model is built once and can then serve any number of queries. Inference is a marginal cost: every query adds computation of its own. As the user base grows and each user asks more of the model, the fixed cost is spread across more use while the marginal cost multiplies. That is why inference is the share that grows, and why it is the measure this paper is focused on.

Each token processed and each training run carries an energy cost, so total energy consumption is pulled in two directions by two different forces. Volume growth pushes it up: more models trained, increased intensity of use, and more users all increase energy consumption. Efficiency, on the other hand, pulls it down. The impact of efficiency gains is greater than most people realise – the energy behind an average token falls by more than 95% by 2040 in our base case. Total demand still rises, because volume grows faster still. The question is not which force wins, but by how much volume outpaces efficiency, and that margin is the thread running through the rest of this paper.

2. Epoch AI, "Data on AI Models", accessed August 2026.



Key terms

Model: The trained generative AI system itself, a large set of parameters learned from data that turns an input into an output. Providers offer families of models sized for different jobs: Claude comes as Haiku, Sonnet, Opus, Fable and Mythos, Gemini as Flash and Pro. Larger models reason better but cost more electricity to run, so the energy behind a request depends on which model answered it.

FLOP (floating-point operation): FLOP measures the total amount of computation performed. This is the standard metric for comparing the scale of training runs across different labs and models. A frontier training run in 2025 uses on the order of 10^{26} FLOP (a 1 followed by 26 zeros).³

Token: A token is the basic unit of text that a large language model (LLM) processes. Roughly, one token equals about three-quarters of a word.⁴ When you type a question into ChatGPT, the system converts your text into tokens, processes them, and generates output tokens as a response. Every token processed consumes a small amount of energy, so the total number of tokens processed globally directly drives electricity demand.

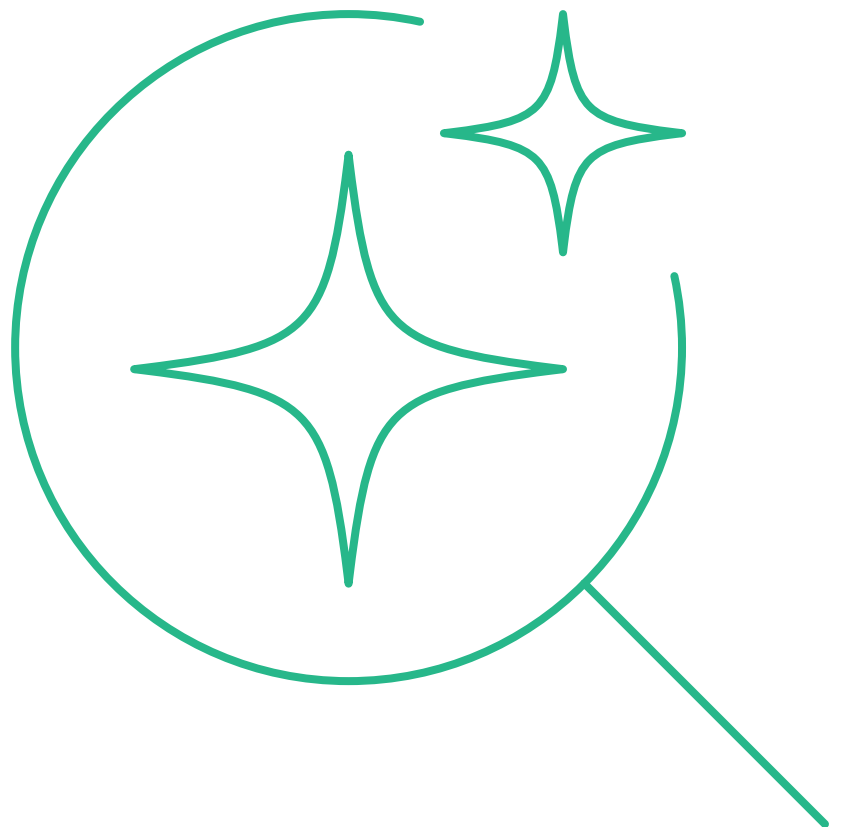
Agentic AI: An agent is a model given a goal and a set of tools, working over many steps rather than answering a single prompt, with each step's output feeding the next. A normal answer is bounded by what a person will read; an agent's is bounded only by time and cost.

3. Epoch AI, AI Models database, accessed May 2026.

4. OpenAI, "What are tokens and how to count them?", accessed August 2026.

SECTION 2

Workload-based projections: Forecasting demand, not supply



Both digital infrastructure and power assets are financed over long time horizons. For anyone building or financing them the question is not how fast data centres go up in the next few years, but how much demand generative AI will place on them over the life of the asset, often the next ten to twenty years.

Most estimates we have seen build projections from supply: power contracted, projects in the pipeline or chips shipped.⁵ Estimates can be constructed via these trackable items because the data exists. But announced pipelines thin out quickly beyond a few years and a chip backlog reaches only as far as the orders. A handful of very large campuses are planned a decade ahead, but most capacity is not, so beyond the near term a projection has to rest on something else. That makes them forecasts of what will be built, not of what will be needed.

That is the gap we have tried to fill. Building from how the technology is used rather than from what is being installed measures the right variable. It also enables the estimate to run as far as the behavioural assumptions can be defended. And by forecasting demand it can be set against the build-out as an independent check on supply rather than a restatement of it.

Two further advantages follow.

- **It isolates generative AI, the segment the capital is chasing, from what headline data-centre figures blend in.** As Section 1 set out, that is where the new capital is concentrated, and where the overbuild question sits. Assessing whether the build-out is justified therefore means looking at generative AI on its own, rather than at a total that would have grown regardless. Our inputs are people, prompts and the output those prompts generate, so cloud computing, storage and other kinds of AI cannot enter the estimate. That distinction will get harder to observe as AI is embedded into ordinary software, but the unit we measure does not change.
- **It enables us to separate inference from training.** The two draw power quite differently and place different demands on generation and the grid, as we discuss in Section 5.

Figure 5 sets out the building blocks. Our analytical framework can be split into two halves, each anchored on what can actually be observed, and added together at the end.

- On the training side, we start from disclosed compute spend at the major laboratories, scale that up to cover the market as a whole, and infer electricity consumption based on the market price of a unit of rented compute.
- On the inference side, we start from how many people use AI and split them by how heavily they use it. We then build up how often each group prompts a model and how much output each prompt generates. That output is split again by the kind of work being done, since a reasoning request costs far more energy than a simple chat.⁶ It is then converted into electricity using published measurements. Images and video are counted separately, per image and per second of clip respectively.

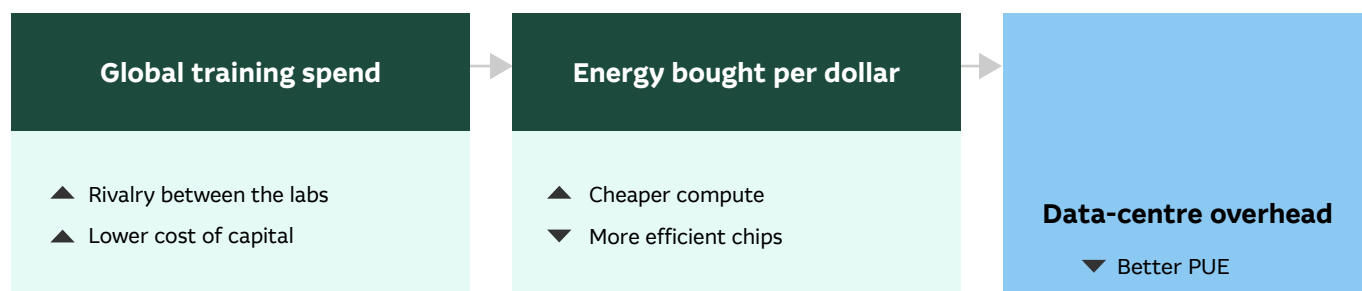
The two are anchored differently because demand reveals itself differently in each. Inference shows up in usage; training is a decision about how much compute to buy, so what reveals it is spending. Neither is a supply measure. Our spend figure is the value of compute consumed within the year, not capital expenditure on hardware.

5. Kamiya, G. & Coroamă, V.C. (EDNA – IEA 4E TCP), "Data Centre Energy Use: Critical Review of Models and Results", March 2025.

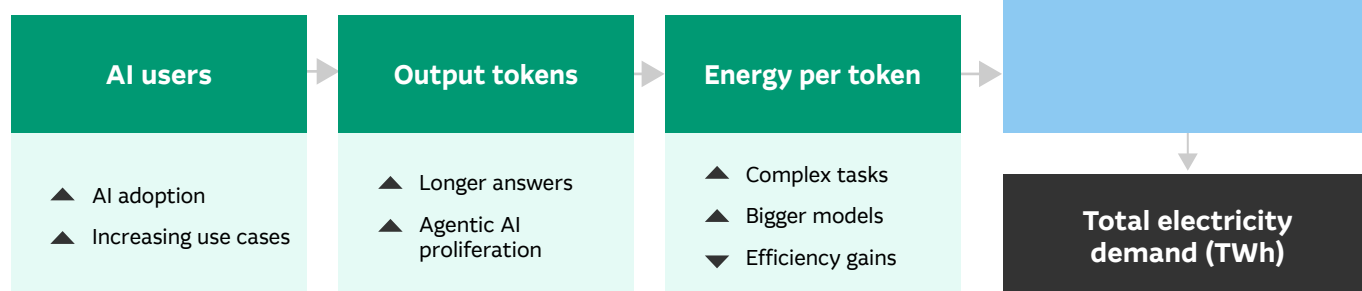
6. Bloomberg, "The rise of AI reasoning models comes with a big energy tradeoff", December 2025.

Figure 5.
How the model builds generative AI electricity demand

Training (anchored on spend)



Inference (anchored on usage)



▲ Raises demand ▼ Lowers demand

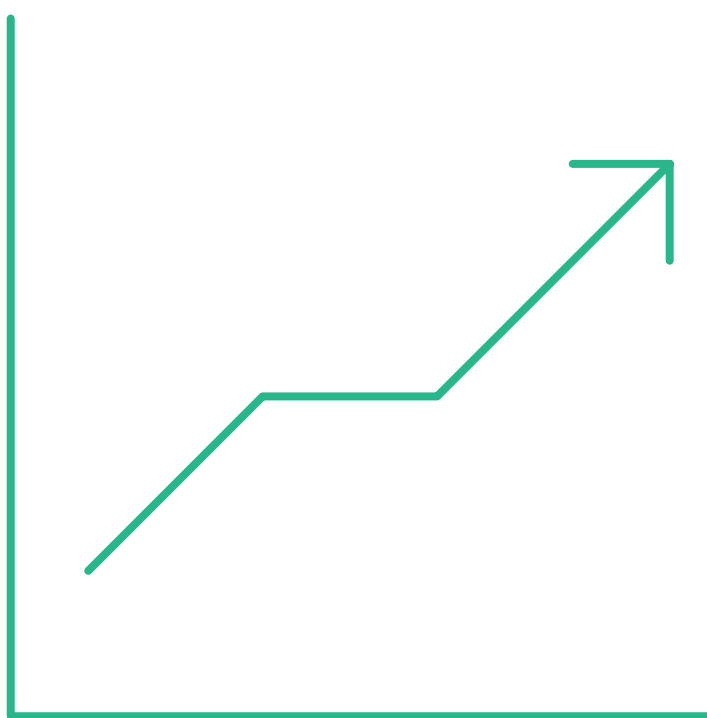
Source: Macquarie Asset Management analysis.

Our analysis runs three scenarios – low, base and high – to span the range of plausible outcomes. Each pairs a coherent set of assumptions about how widely AI is adopted, how intensively it is used, how quickly efficiency improves, and how fast training spend grows. They are best read as three internally consistent cases rather than as statistical confidence bounds around a central estimate. The headline figures throughout this paper refer to the base case. Full mechanics are in the Appendix.

- **The base case: More people using AI, and each using it far more.** Adoption reaches around 70% of the world by 2040, comparable to global internet penetration today, helped by generative AI being absorbed into search, office software, and websites, so that users are reached without choosing to adopt anything. Requests get longer and harder, with output lengths rising nearly sixfold, and more of that work runs through agents, where the output is fed straight back in as the next step rather than read by a person. Training spend follows currently published forecasts through the rest of this decade, then slows as the build-out matures. Efficiency keeps improving, but the pace slows as chips approach their physical limits and the easiest algorithmic gains are exhausted. By 2040 the base case projects the volume of work rising around 400-fold, and electricity demand around ninefold.
- **The low case: Adoption stalls, and use stays light.** This scenario describes a bounded role for AI. Adoption reaches only 55% of the world by 2040, well short of internet penetration today. Usage intensity stays modest and agentic use spreads more slowly than in the base case. Training spend plateaus by the late 2030s. Efficiency also improves more slowly, on the view that a smaller market draws less investment in chips and software alike. AI proves useful but never ubiquitous: by 2040 the volume of work rises around a hundredfold and electricity demand around fivefold.
- **The high case: AI reaches nearly everyone, and machines do the asking.** This scenario assumes pervasive agentic use, with AI reaching around 90% of the world's population by 2040. Outputs lengthen sharply as they are written for machines rather than people. Training spend roughly triples by 2030. Efficiency improves faster here than in either other case but is still comfortably outrun by volume: by 2040 the volume of work rises more than a thousandfold and electricity demand thirteenfold.

SECTION 3

Key results: Electricity demand rises ninefold



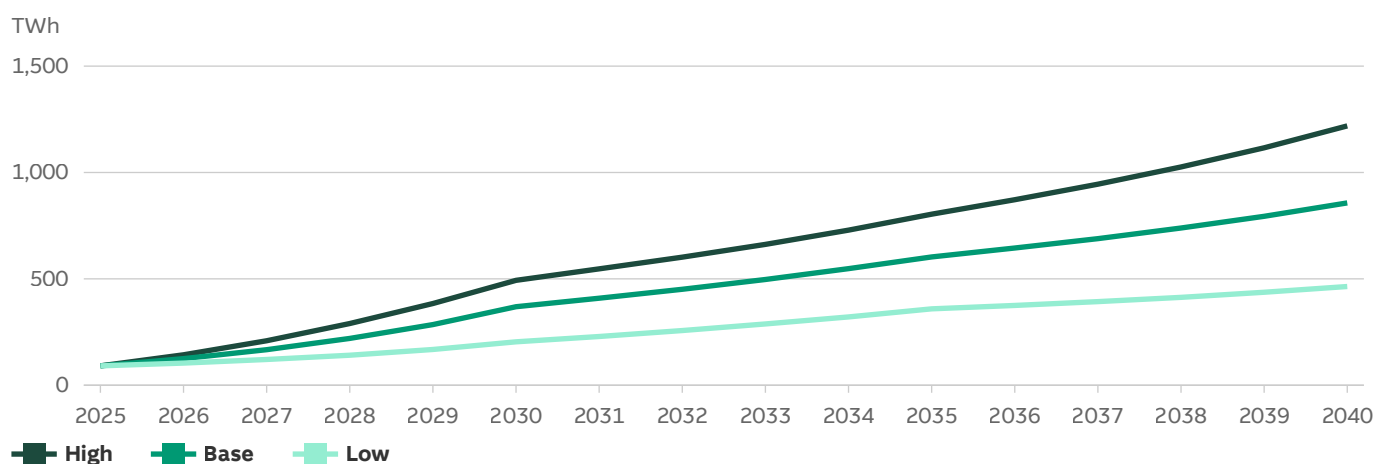
Based on our assumptions, generative AI's electricity use rises from about 91 TWh in 2025 to roughly 857 TWh by 2040 in the base case, a more-than-ninefold increase or about 16% a year (Figure 6).

The scenarios fan out around it: even in the low case demand increases by roughly fivefold to about 460 TWh, while in the high case it reaches some 1,219 TWh, thirteen times today's level.

The wide range reflects the very real uncertainty around adoption, usage intensity and efficiency fifteen years out. It bounds the ceiling, and the ceiling is an interesting number. In the high case generative AI reaches about 3% of global electricity in 2040, using more electricity than Germany and France combined do today. That is sizeable, and will require large sums of investment in generation and grid networks to achieve. In the base case it is 2.2% and the low case 1.2%.

The range should be read as an envelope rather than a confidence interval. Each scenario bundles adoption, usage intensity and efficiency into one coherent world and the inference and training halves are then paired scenario for scenario. The high case assumes the fastest efficiency gains as well as the heaviest use. A true upper bound, pairing the heaviest use with the slowest efficiency, would sit above our high case. We do not model it, however, on the view that the same investment that drives heavy use also funds the efficiency gains, so the two should not move in opposite directions.

Figure 6.
Electricity demand in different scenarios



Scenario	2025	2030	2035	2040	CAGR 25-40
Electricity demand (TWh/yr)					
Low	91	204	359	464	11.5%
Base	91	369	603	857	16.1%
High	91	493	804	1,219	18.9%
Average continuous draw (GW)					
Low	10	23	41	53	11.5%
Base	10	42	69	98	16.1%
High	10	56	92	139	18.9%
Generative AI as % of global electricity					
Low	0.3%	0.7%	1.0%	1.2%	
Base	0.3%	1.2%	1.7%	2.2%	
High	0.3%	1.6%	2.3%	3.1%	

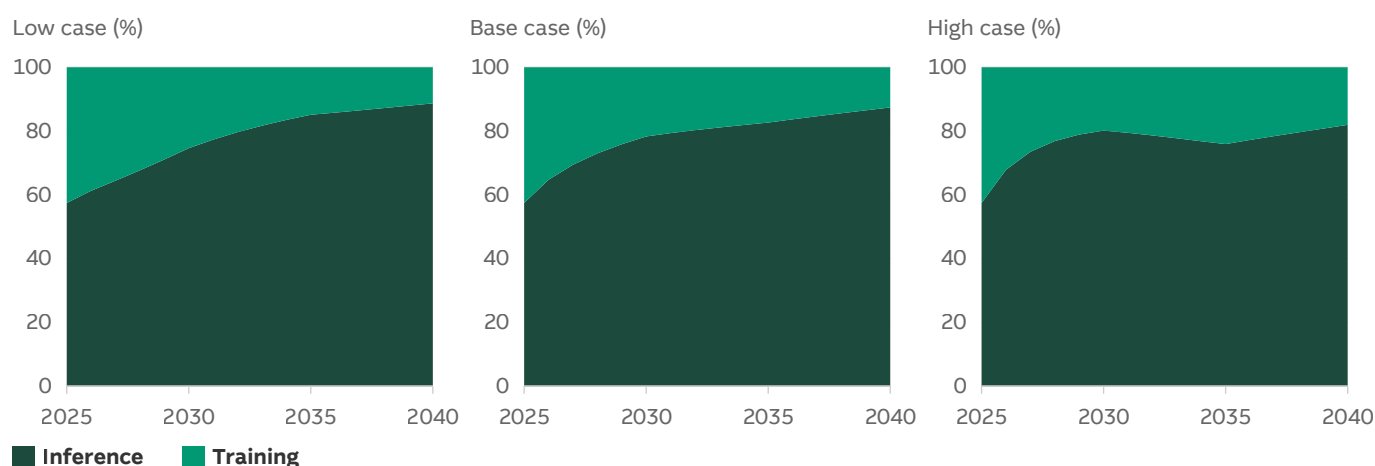
Source: Macquarie Asset Management analysis. Global electricity denominator: BloombergNEF (BNEF) New Energy Outlook 2026, Economic Transition Scenario, final consumption basis. Implied average continuous draw (GW) is annual TWh divided by the hours in a year (8,760 hours).

Three features of the projected demand composition are worth drawing out:

- **Inference dominates and increasingly so.** Inference already accounts for more than half of demand today, and it is projected to rise to 87% by 2040 in the base case. Training is a fixed cost while inference is a marginal one that recurs with every request, so the balance tips further towards inference as usage grows. That holds across all three scenarios, which land between 82% and 89% by 2040. Training retains the largest share in the high case, where the frontier build-out keeps compounding even as usage surges (Figure 7).

Figure 7.

Workload mix, inference versus training

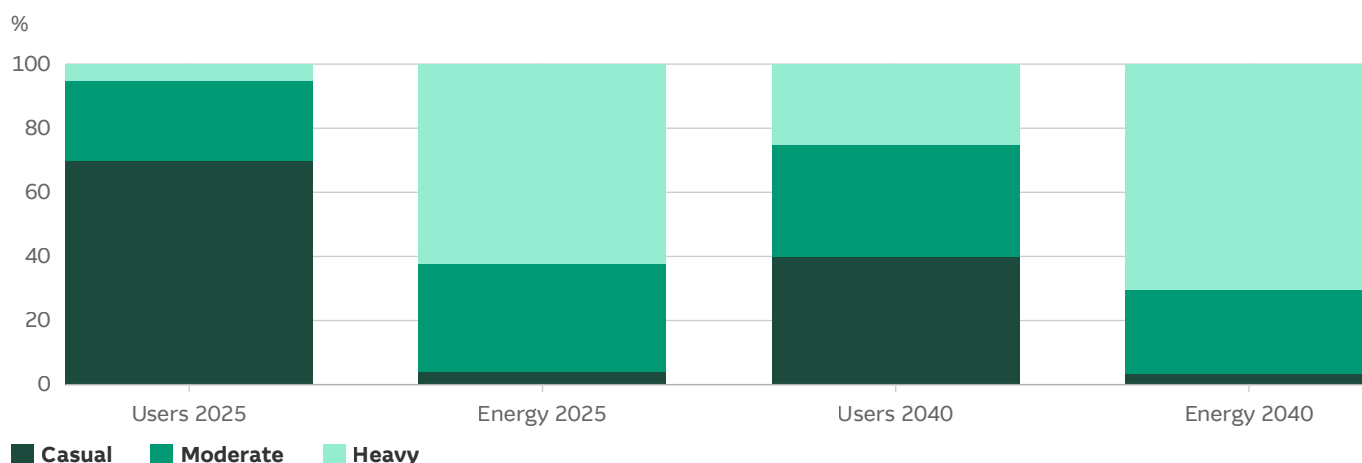


Source: Macquarie Asset Management analysis.

- **A small population carries most of the inference workload.** We split users into three groups: casual users making occasional short requests, moderate users treating generative AI as a routine work tool, and heavy users running it through an API or in agentic loops. The heavy group is 5% of users in 2025 but accounts for just over 60% of inference energy; by 2040 it is a quarter of users and 70% of the energy (Figure 8). A heavy user draws more than two hundred times the electricity of a casual one today, narrowing to around thirty times by 2040 as use intensifies across the board. Casual users fall from 70% of the modelled population to 40%, and account for just over 3% of energy consumed by 2040. Volumes like these are only affordable because the price of a token falls with the energy behind it. At today's prices a heavy user would be spending well over \$US100,000 a year by 2040, against something closer to a software licence on our efficiency path.
- **Video stays small.** Video generation attracts attention because a single clip is expensive – around 200 watt-hours (Wh) for six seconds at 2025 rates including data-centre overheads, enough to charge a dozen smartphones. But it is 2.5% of demand in our base case and under 5% even in the high case, which assumes some 2 billion minutes of video generated every day by 2040. It is the most uncertain part of our estimate; we take a closer look in Section 4, where we found that even pushing every video assumption against us at once, which nothing today supports, leaves it at around a fifth of demand in our base case.

Figure 8.

Share of users against share of inference energy in our base case



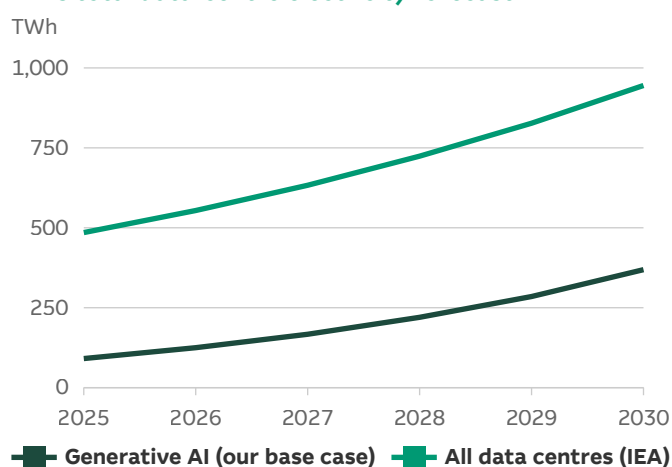
Source: Macquarie Asset Management analysis.

Comparing our figures with published forecasts

Very few forecasters project generative AI on its own, and we have found no projection of it to 2040. Published estimates are drawn wider: some cover all AI data centres, others all data centres, with generative AI one component of each. Of what is available, the most useful test is what our figures imply about generative AI's share of the data-centre estate, set against what others assume.

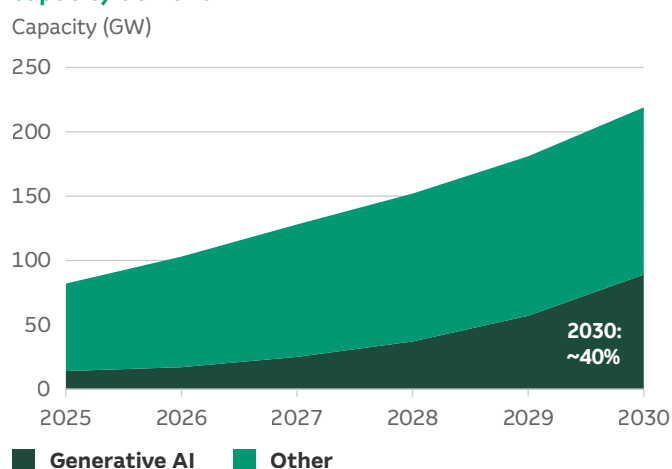
Against the IEA's forecast for all data centres, our base case suggests 39% by 2030 (Figure 9). McKinsey is the only major forecaster we have found that splits generative AI out separately, and its projection has generative AI at around 40% of total data-centre capacity demand by 2030 (Figure 10). The two are not strictly like for like, since McKinsey measures data-centre capacity in gigawatts rather than electricity demand in TWh, but we both arrive at the same share despite being built on entirely different foundations: ours from user behaviour upward, McKinsey's from capacity downward.

Figure 9.
Our base case for generative AI demand against the IEA's total data-centre electricity forecast



Source: Macquarie Asset Management analysis; IEA, "Key Questions on Energy and AI", April 2026.

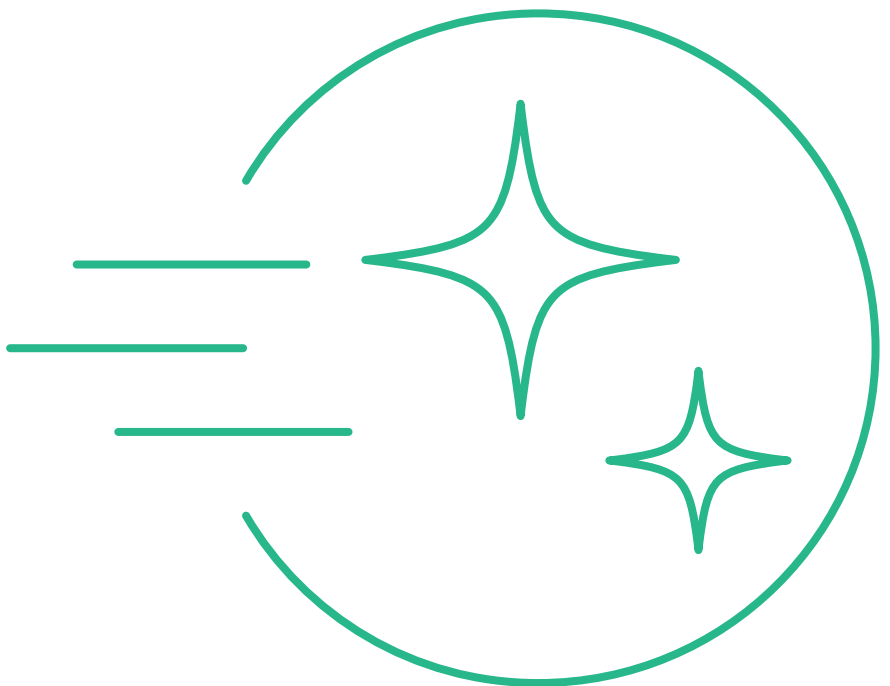
Figure 10.
McKinsey's generative AI share of data-centre capacity demand



Source: McKinsey & Company, "AI power: Expanding data center capacity to meet growing demand", October 2024.

SECTION 4

The main levers: Efficiency versus intensity, and the video wild card



Our analysis rests on a range of inputs, from how widely and quickly AI spreads to how efficiently each token is served. Figure 11 shows how varying each of the main inputs, holding everything else constant, would change our base-case estimate of 2040 electricity demand.

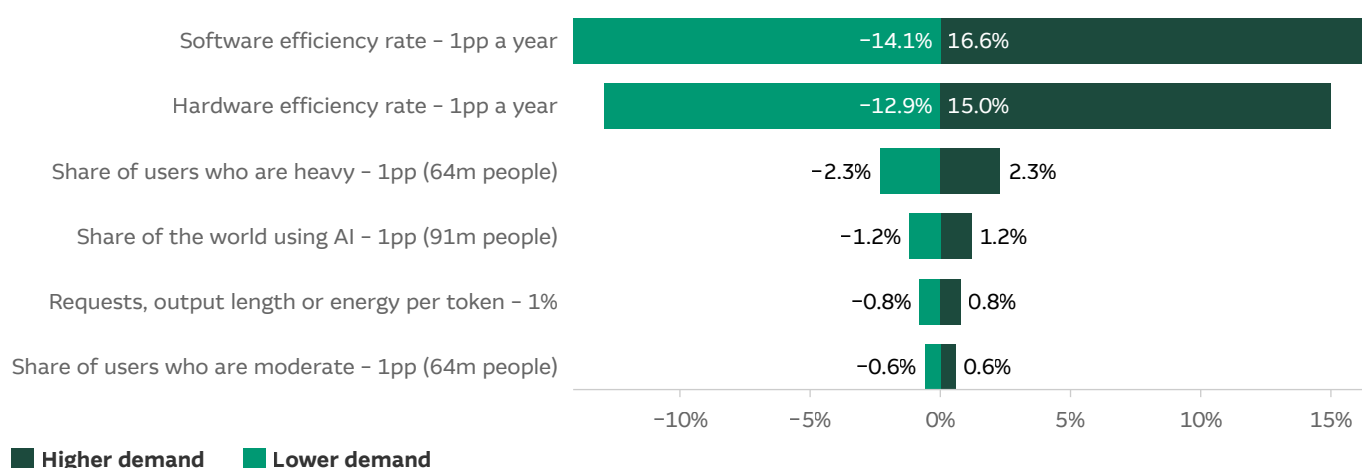
Three observations stand out.

- **Who uses AI matters far more than how many use it.** Shifting 1% of generative AI users from casual to heavy, some 64 million people, moves 2040 demand by 2.3%. Bringing 1% of the world's population online as new users, without changing the user mix, moves it by only 1.2%.
- **The token drivers behave alike, because the arithmetic makes them.** Inference is the product of requests, output length and energy per token, so a 1% change in any one of them has the same effect – it moves demand by the same 0.8%. What separates them in practice is not leverage but how far each is likely to move.
- **Efficiency has the biggest impact on demand.** One percentage point a year on either rate, hardware or software, moves 2040 demand by 13% to 17% – several times any user or volume driver here. A single point sounds small, but it applies every year for fifteen years, and the difference accumulates.

The rest of this section takes them in turn: the race between efficiency and volume that sets the level; the intensity of use that separates the scenarios; the one part of the picture that could still surprise us; and why training is bounded by capital rather than use.

Figure 11.

Sensitivity analysis – how far each assumption moves the 2040 estimate



Source: Macquarie Asset Management analysis.

The level of power demand: Efficiency versus volume

The level of demand is set by a race between the rising volume of computation and the falling energy cost of each unit of it. That the volume will rise is close to certain. The open question is whether efficiency falls quickly enough to hold the total in check.

Two forces push the energy cost of a token down:

- **Hardware.** Each generation of AI chip performs more computation per watt than the last. On Epoch AI's data, the energy efficiency of leading machine-learning hardware has improved by a factor of roughly 1.3x per year over the past decade, though the pace is easing as chip design approaches physical limits (Figure 12).
- **Software.** Better model architectures, together with techniques such as quantisation,⁷ distillation⁸ and the use of smaller task-specific models steadily cut the computation needed for a given result. Our base case assumes software efficiency gains arriving at roughly 25% a year today, decaying to 10% by 2040 as the stack matures and algorithmic headroom thins.⁹

7. Quantisation reduces the numerical precision a model uses, so each calculation is cheaper without much loss of quality.

8. Distillation trains a smaller model to reproduce the behaviour of a larger one, giving most of the capability at a fraction of the running cost.

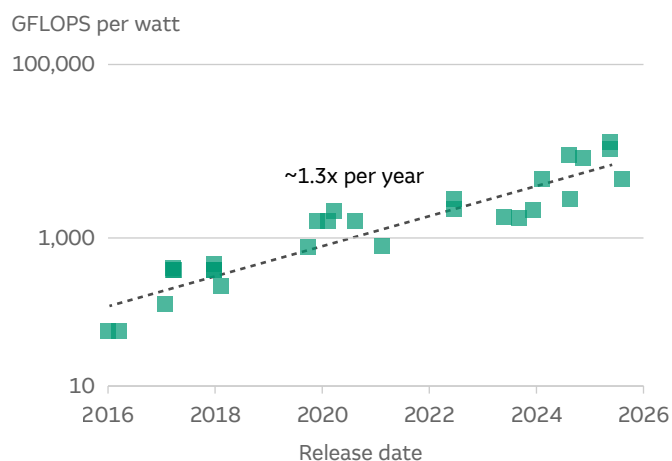
9. A 2026 Sakana AI and NVIDIA collaboration gives a sense of scale: a single advance, cutting out computation the model was doing but not using, delivered roughly 20 to 30% more output from the same chips. Source: Cetin, Peluchetti, Castillo, Naruse, Murakami and Jones, "Sparsier, Faster, Lighter Transformer Language Models," arXiv:2603.23198, ICML 2026.

Set against these gains, model growth pushes back. Frontier models have grown to hundreds of billions of parameters up from millions just a decade ago. Greater capability generally means more computation per token: if a chip is twice as efficient but the model running on it is twice as large, the energy behind a token has not changed. Together with a shift towards more demanding tasks such as reasoning, that absorbs much of the gain. Hardware and software improvements alone would cut the energy behind each output token around 125-fold by 2040; the net fall is around 21-fold.

A further improvement happens outside the chip, as data-centre overheads fall with better cooling, better power usage effectiveness (PUE)¹⁰, and higher utilisation. Those gains narrow as PUE approaches its practical floor and utilisation approaches its ceiling, but our base case still has electricity per unit of AI output falling around 28-fold by 2040. Google's disclosure that its median Gemini prompt now draws 0.24 watt-hours, 33 times less than a year earlier, is the same story at the level of a single query.¹¹

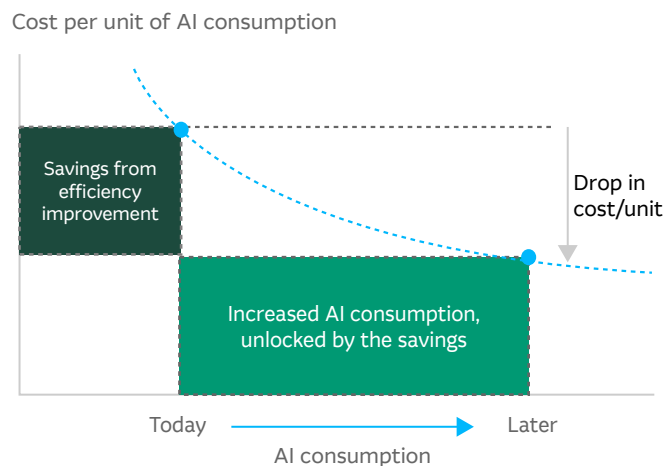
Yet none of this is enough to offset volume growth. Over the same fifteen years the volume of tokens rises 399-fold, against a 28-fold fall in the electricity behind each one. That leaves a fourteenfold rise in the power behind text inference, and a ninefold rise in total demand once training and multimodal content (image and video) are included. Efficiency does the great majority of the work here, but cheaper output invites more of it, and volume grows faster than efficiency can absorb. This is the Jevons paradox¹² in arithmetic (Figure 13).

Figure 12.
Energy efficiency of leading AI hardware, 2016–2025



Source: Epoch AI.

Figure 13.
The Jevons paradox



Source: Macquarie Asset Management.

10. Power usage effectiveness: total data-centre electricity divided by the electricity reaching the computing equipment, where 1.0 would mean nothing at all spent on cooling or power conversion.

11. Elsworth et al., "Measuring the environmental impact of delivering AI at Google Scale", arXiv:2508.15734, August 2025.

12. Named after the economist William Stanley Jevons, who observed in 1865 that more efficient steam engines increased rather than reduced Britain's coal consumption, because cheaper power made more uses of it worthwhile. The same pattern has since been observed in lighting, computing and data storage.

Volume: Determined by agentic AI

The volume that efficiency has to race against depends less on how many people use AI than on how intensively each of them does. The gap between a light user and a heavy one runs to more than 200-fold today, narrowing only gradually to about thirtyfold by 2040. Efficiency determines how far the energy cost of each unit falls.

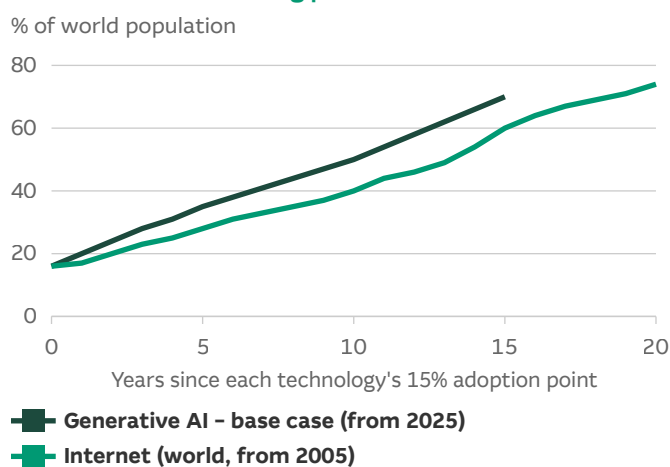
Our base case generative AI adoption path runs ahead of the internet's. Indexed to the year each technology reached around 15% of world population, it stays above the internet's own path throughout, reaching 70% of the world at year fifteen where the internet reached 60% (Figure 14). That is consistent with what we have seen so far: generative AI has diffused faster than almost any technology in history, and adoption is increasingly passive, embedded into search, office software and websites so that users are reached without choosing to adopt anything.¹³

Even so, adoption is not where the majority of the growth comes from. Of the 399-fold rise in tokens from 2025 to 2040, a wider user base contributes a factor of 4.8x. The other 83x comes from each user generating more (Figure 15). Reasoning models write longer answers and agentic systems consume their own output as the next input.

That last mechanism removes a limit that used to hold. By 2040 a heavy user in our base case generates around 3.9 million output tokens a day – some 190 hours of reading. These are not volumes a person consumes. The heavy segment is really a measure of agentic activity, which is why per-user intensity grows 83-fold while the user base grows less than fivefold. Indeed, we are already seeing the first clear sign of it: Microsoft reports 2.3 million GitHub pull requests associated with AI agents in March 2026, a 28-fold rise in ten months.¹⁴

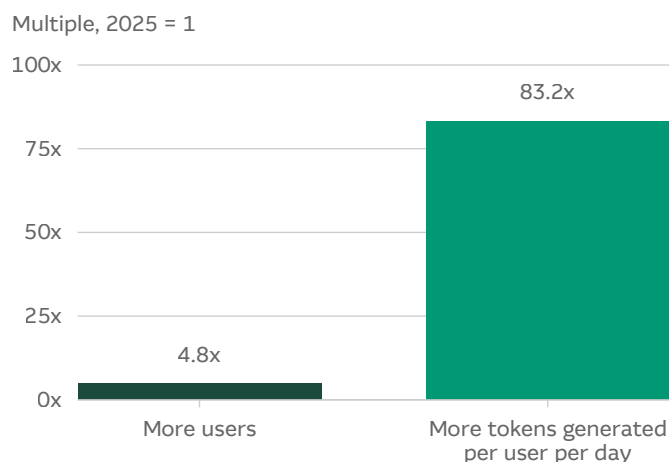
Ours is therefore an intensity range. The question for an investor is not whether AI is widely adopted, but how demanding each user turns out to be.

Figure 14.
Generative AI adoption against the internet,
from a common starting point



Source: Our World in Data; Macquarie Asset Management analysis.

Figure 15.
Where token growth comes from, base case 2025–2040



Source: Macquarie Asset Management analysis.

13. Epoch AI, "After the ChatGPT moment: Measuring AI's adoption", July 2025.

14. Microsoft, "Global AI Diffusion Q1 2026 Trends and Insights", accessed August 2026.

Video: The wild card

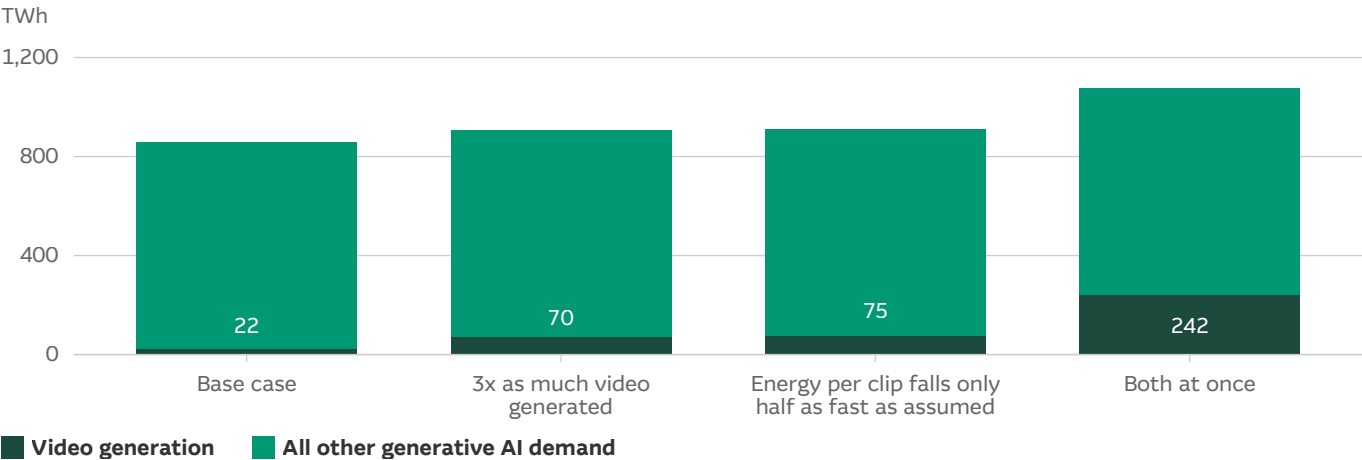
Video is where our estimate could most easily be wrong. The per-second energy behind a generated clip is the least settled measurement we use, with published estimates of the energy behind a second of generated video differing by an order of magnitude.¹⁵

We estimate video accounts for about 2% of inference electricity today, despite being far more energy intensive per unit than text. A six-second clip draws roughly 34 times the energy of a text query. But for every clip generated there are some 2,700 text queries, and the gap is structural: a text response is produced for one person and used once, while a video is produced once and watched by many. Video generation scales with how much is made, not with how much is watched.

Within the range we can defend, the numbers stay small. Our high case assumes around 2 billion minutes of video generated a day by 2040. Stress-testing the two assumptions that matter most does not change the conclusion. Generating three times as much video takes it from 22 to 70 TWh, and halving the rate at which per-clip energy falls takes it to 75 TWh, in both cases under a tenth of total demand (Figure 16). Only if both hold at once does video become material, at 242 TWh or a fifth of the total, and nothing today supports that combination.

Two things could change this picture. The first is AI displacing filmed content, and it is already happening in some pockets: in China, more than 95% of the 128,000 micro-dramas¹⁶ launched in the first quarter of 2026 were AI-generated.¹⁷ But this is bounded, for the reason above: content made once and watched by many does not need much generating. The second would not be bounded. If video came to be generated for each viewer rather than made once and shared, generation would scale with how much is watched rather than how much is made, and the numbers above would be orders of magnitude too low. Nothing we can point to suggests that is imminent. It is also beyond what we can model, and we have not attempted it.

Figure 16.
Stress-testing AI video demand



Source: Macquarie Asset Management analysis.

15. ML.ENERGY, "Diagnosing Inference Energy Consumption with the ML.ENERGY Leaderboard v3.0", January 2026.
16. Micro-dramas are vertical-format serials made for phones, typically one to three minutes an episode across sixty to a hundred episodes.
17. 36Kr, "AI Short Drama Survivor Game", August 2026.

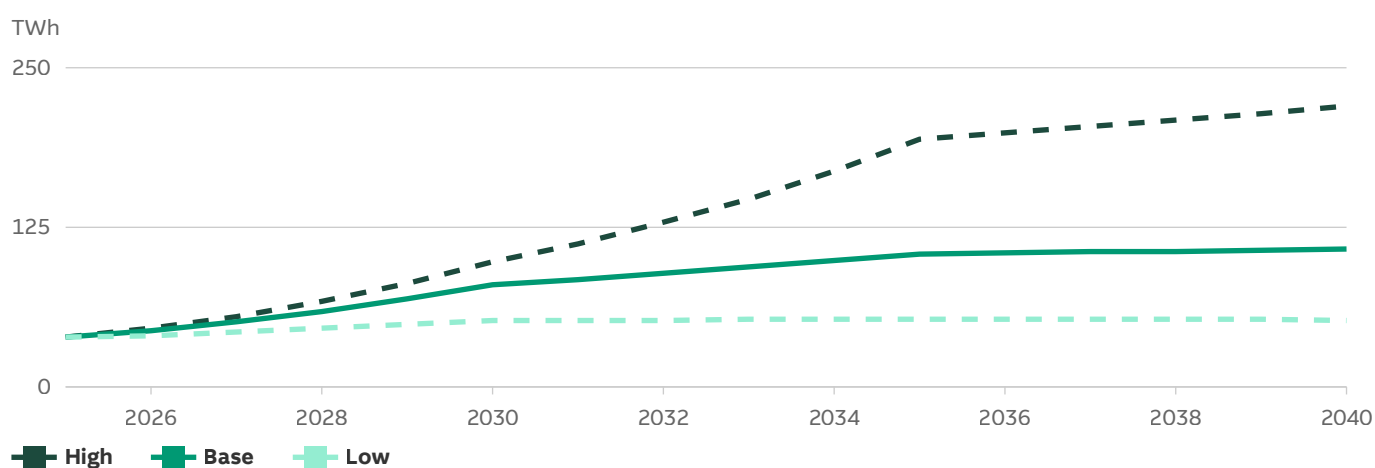
Training is bounded by capital, not by use

Training is a deliberate choice by laboratories about how much to spend, so that is where we start. We estimate global spend on compute based on what the leading laboratories disclose, then convert that spending into electricity.

What varies between our cases is how fast that spending grows. The base case has it growing at about 20% a year to 2030, the low case at half that rate and the high case at a quarter above it. All three slow thereafter, on the view that the current phase is a build-out rather than a steady state, easing to 3% a year in the base case by the late 2030s. That takes training electricity from 39 TWh in 2025 to between 52 and 220 TWh by 2040 (Figure 17).

We do not model training from compute growth, which would mean extrapolating the historical trend. Compute growth is not really a trend at all. It is a choice the labs make each year, guided by how much capital they can raise and justify. The trend is also unlikely to hold to 2040, since continuing it would run into constraints on power, chip manufacturing and data well before then.¹⁸ Modelling the spending directly is the more practical route.

Figure 17.
Training electricity demand in different model scenarios

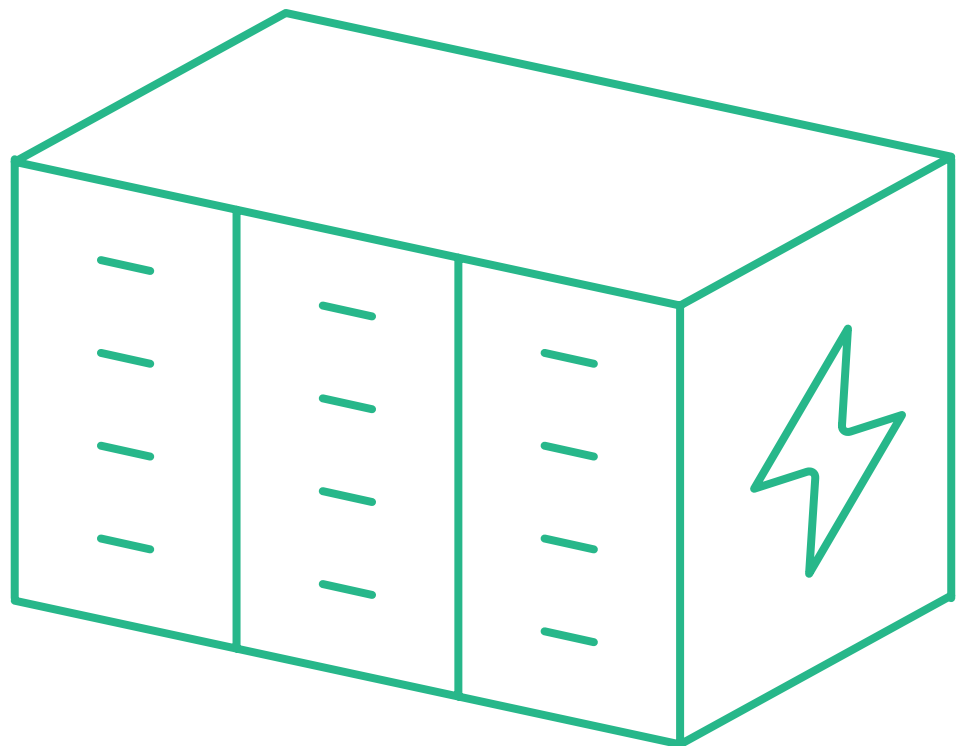


Source: Macquarie Asset Management analysis.

18. Sevilla et al., "Can AI scaling continue through 2030?", Epoch AI, August 2024.

SECTION 5

Implications for investors: The scale of the build and the kind of load it serves



The implications for infrastructure are about how much power is needed, but even more about where, when and how steadily. The world can generate what generative AI will need. In our base case it reaches about 2.2% of global electricity by 2040.

That might seem like a modest share, but it sits on top of the demand growth already coming from electrification, industry and heating, and it arrives faster than any of them. Getting that power to the right places, at the moments it is needed, is the harder problem. It also becomes more difficult as the load grows, because the part of it that can wait is shrinking. Meanwhile the growth is increasingly driven by machines rather than people, which changes not only how much electricity is needed but when.

What this means in investment requirement terms

Our base case implies around 120 GW of installed data-centre capacity by 2040, some 106 GW more than today. This will cost roughly \$US4.5 to \$US5.6 trillion to build, or \$US300 to \$US375 billion a year.¹⁹ About 40% of this is the data centre itself plus the generation to power it and the grid to connect it: land, shell, cooling plant, electrical systems, substations, generating capacity and transmission. The other 60% is IT equipment such as servers and networking. Our figure is the cost of the capacity added by 2040 in today's cost terms and takes no account of IT equipment turning over during the period, so it likely understates the total investment requirement.

Demand is large and durable but not a threat to global supply

On our base case, generative AI reaches about 2.2% of world electricity consumption by 2040, and 3.1% at the top of our range, measured against a global supply that grows substantially over the same period. That is a large load in absolute terms, but it is achievable if sufficient capital can be mobilised.

A natural question is whether efficiency gains simply offset the growth in volume. They do not. Without the improvement we assume, 2040 demand in our base case would be around 230 times today's. Even with it, demand rises ninefold. Cheaper output has consistently stimulated more demand, so the energy behind a single query has shrunk while total consumption has climbed. We have seen this before in cloud storage: the cost of storing a gigabyte fell by roughly 87% between 2009 and 2022, and the world responded by storing six times as much.²⁰ Moreover, even if demand disappoints, it does not disappear. In our low case it still rises around fivefold by 2040, to about 464 TWh, growing every year of the period. It describes a world where AI use stays light, not one where the load goes away.

19. Our own calculations. Capacity is measured by the power the IT equipment draws, so a gigawatt of capacity means a gigawatt of servers plus the building, cooling and electrical systems sized around them. Derived from our demand estimate using the utilisation and overhead paths in our analysis. Figures are the cost of the capacity added by 2040 at today's prices and exclude IT equipment turnovers.

20. Backblaze, "Hard Drive Cost Per Gigabyte", November 2022.

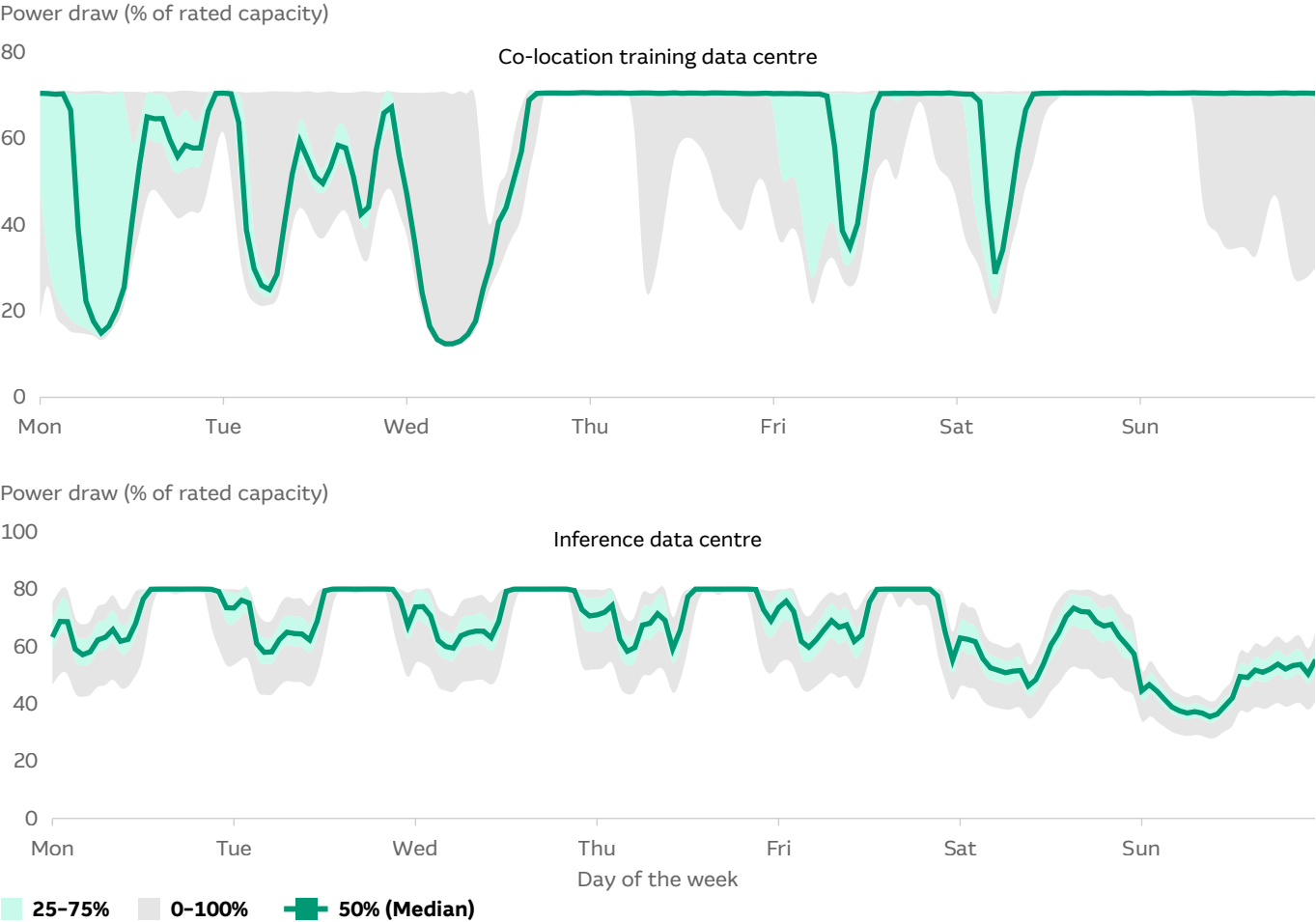
Not all AI load is created equal and the difference matters for the grid

Generation adequacy assumes power can get where it is needed, when it is needed. In practice it must be built in specific places, connected through networks that take years to permit and build, and delivered at the moment the load asks for it.

All three of those factors depend on what kind of load it is, a point ignored in much of the literature on AI and its power demand implications. Training is patient but erratic. It can wait and doesn't need to be done close to its users, but its draw swings from hour to hour with little regular pattern. Inference is demanding but more predictable. It must answer the moment it is asked, yet its load follows a steady daily and weekly rhythm, troughing before dawn, peaking in the late afternoon, easing at the weekend (Figure 18).

That is why the split matters. Published estimates cover all AI, or all data centres; they rarely separate generative AI from other AI workloads, let alone training from inference within it. Ours does both, and shows inference rising from 57% of demand today to around 87% by 2040 in our base case. For a network operator or a site developer, what matters is the training and inference split rather than whether the tokens came from a chatbot or from a spreadsheet that quietly called a model. Splitting the estimate is what makes the discussion possible.

Figure 18.
Training swings, inference repeats: Facility power profiles at 80% average utilisation, by hour of the week



Shaded bands show the range of simulated power outcomes over one year. Wider bands indicate greater uncertainty in demand at that time of the week.

Source: Vercellino, R. et al., “Measurement of Generative AI Workload Power Profiles for Whole-Facility Data Center Infrastructure Planning”, National Laboratory of the Rockies, April 2026. Simulated profiles for a 10 MW co-location site running training and fine-tuning workloads and a 1 MW inference site, each shown as a share of its rated capacity. The training profile is net of queueing; jobs are scheduled first-in-first-out and at this utilisation around three-quarters wait 6.5 hours, so what remains reflects when work arrives rather than an absence of scheduling.

Furthermore, there may also be a geographic dimension to this. Training concentrates where the model developers are, which today means overwhelmingly the United States, with China running a parallel ecosystem of its own. A frontier run needs one very large site, and it will be built close to the firm doing the training rather than close to anyone using the result. Inference does not work that way. It follows the users, and increasingly the rules governing where their data may be processed, as sovereignty requirements and sector regulation push serving capacity into the markets it serves. As inference comes to dominate the load, the direction of travel is toward a more geographically dispersed one, with more of it landing in Europe and other regulated markets rather than in the few countries where training clusters today.

Agentic use could break inference's link with human behaviour

The load shape for the inference data centre in Figure 18 is one driven by human behaviour. It tracks the working week because a person is waiting at the other end of each request. Agentic use would likely sever that link as it proliferates, and would do so in two ways at once:

- **The first is level.** Demand stops being bounded by how much a person can read or type, which is what allows per-user intensity in our analysis to grow 83-fold while the user base grows less than fivefold.
- **The second is shape.** Work that runs unattended has no reason to follow the working day. In principle that flattens the profile and raises utilisation, which suits a grid: a flatter load needs less capacity behind the same energy. In practice a price signal would do much of that work on its own, since load chasing the cheapest hour bids that hour up until it is no longer the cheapest.

Generative AI providers already discount work that can wait, and some are beginning to shift it around grid conditions, but these arrangements are at an early stage.²¹ Not all work can wait, either. Many users would rather pay for an answer now, and some requests lose their value if delayed. Where the load cannot be steered, flexibility can be built around it instead: on-site generation, storage, or a connection sized for the swings rather than the average. That is familiar ground for infrastructure investors, and easier than waiting for AI pricing to follow the grid.

21. OpenAI and Anthropic both halve the price of requests that can be returned within 24 hours; see OpenAI, "Batch API". Google has written around 1 GW of demand response into contracts with five US utilities, shifting or curtailing machine-learning workloads when the grid is under stress; see Google, "Google signed 1 GW of data center demand response", March 2026.

Appendix: Generative AI electricity demand – model methodology



We built a proprietary model, which forecasts annual electricity consumption from generative AI and agentic workloads worldwide to 2040. The model is primarily conceived as two independently anchored halves, inference (serving models) and training (building them), as an attempt to address electricity drawn from the grid, including the cooling and power losses of the data centre around the chips.

1. Scope

We focus our work on generative and agentic AI only – text, image and video generation, plus the full training cycle including pre-training, fine-tuning, reinforcement learning and post-training. It excludes everything else running on AI accelerators: recommendation and ranking, advertising models, computer vision, speech, classical machine learning at scale and GPU-based high-performance computing.

Given our narrower approach, it is plausible readers might find our numbers differ from published headline “AI electricity” estimates out there. It is not practical to review all of these published estimates. In preparing this work, however, we found that practitioners define “AI electricity” in a wide variety of ways, on both the consumption and the supply sides. That topic probably warrants a separate discussion in itself. Our scope is rather limited in this regard to the activities mentioned, which, as explained in the main body, we have chosen to focus on for their commercial relevance and high impact. We also note the need for further work in other aspects of “AI” in the industry that is beyond the remit of this paper.

2. Modelling – the two halves are anchored differently

To conceive our approach simply, we see inference as bound by how much work people ask the models to do: put simply, how many people use AI, how often, and how much output each request generates. Training is the other half and we approach it slightly differently. To us it appears bound by the money available to spend on compute. Each half is anchored on industry disclosure and informed by academic measurement. We believe there is rather limited overlap between the two areas, and simple mathematical sums might suffice to give a high-level view.

The asymmetry of the two anchors matters. A work-based anchor carries no price, so efficiency gains must be applied to it directly, whereas a spend-based anchor already embeds them in the falling cost of compute. Applying an efficiency improvement to both halves would raise the question of double counting and whether we can precisely delineate the interaction between the two.

2.1 Inference: from users to electricity

We estimate a user base simply via population multiplied by the share of it using AI, split across three segments by how intensively they use it: casual, moderate and heavy. Cultural intricacies and geographical realities obviously will be factors that are too complicated for us to fully scope in the current remit of our work. To help ourselves conceptualise, we imagine a casual user might be represented by a delivery driver using generative AI in private life, a moderate user an accountant or lawyer for whom it is a routine work tool, and a heavy user a developer calling an API, a quant researcher running research workflows, or any professional whose work now runs continuously through an LLM.

$$\text{AI Users}(\text{year}) = \text{Population}(\text{year}) \times \text{AI Diffusion}(\text{year}) \times \text{Segment Mix}(\text{segment}, \text{year})$$

We observe that the mix of AI usage shifts toward the heavier segments over the horizon. As models become more capable, more work can be delegated to them. As generative AI loses its novelty, using it becomes part of how work gets done rather than something people try. Our model anticipates users move up the segments as a growing share of what they do runs through a model.

Inference volume is measured in output tokens, the text a model generates. Each segment has its own query frequency and output length, both rising over time.

$$\text{Output Tokens}(\text{year}) = \sum \text{segments} [\text{Users} \times \text{Queries per User per Day} \times 365 \times \text{Output Tokens per Query}]$$

We see tokens split across five task archetypes: chat, reasoning, image understanding, video understanding and code completion. The same token costs materially different amounts of energy depending on the work behind it. As reasoning grows as a share of the total, it is reasonable to expect average intensity rises on top of the growth in volume.

In our approach to modelling energy levels, we refer to Jegham et al. (2025), which include the energy a model uses to read a question as well as to write the answer.²²

$$\text{Energy per Output Token} = \text{Base Energy per Token} \times \text{Task Multiple} \div \text{Efficiency Index}$$

The efficiency index compounds three annual rates: hardware gains, software gains, and an offset for models growing larger.

$$\text{Efficiency Index}(\text{year}) = \prod \text{years} [(1 - \text{Hardware Gain}) \times (1 - \text{Software Gain}) \times (1 + \text{Model Growth})]$$

The steps above give the energy used while chips are doing useful work. The electricity a data centre actually draws is higher, for two reasons: chips spend some of their time switched on but idle, and the building itself uses power for cooling and power conversion, measured by power usage effectiveness (PUE).

$$\text{Electricity} = \text{Compute Energy} \times \text{Idle Adjustment} \times \text{PUE}$$

For image and video generation, we model energy per image and per second of video. Consumer volumes grow with the user base and enterprise volumes on their own path. Energy per second of video shares the hardware and software gains used for text, with an offset for rising resolution, frame rate and clip length.

$$\text{Video Electricity} = \text{Videos per Day} \times 365 \times \text{Clip Length (s)} \times \text{Wh per Video-second} \times \text{Overhead}$$

2.2 Training: from spend to electricity

Our approach to modelling training demand is informed by the compute spend disclosed by the large pure-play generative AI laboratories, scaled up to a global figure by an assumed coverage share. Using pure-play laboratories keeps non-generative training out of scope.

$$\text{Global Training Compute Spend} = \text{Observed Laboratory Compute Spend} \div \text{Coverage Share}$$

$$\text{Training Electricity} = \text{Global Training Compute Spend} \times \text{Energy per Dollar} \times \text{PUE}$$

Energy per US dollar is the electricity a dollar of computing buys: the power a chip and its server draw, divided by what an hour of that chip costs to rent. We hold it constant, assuming chips get cheaper to rent at roughly the rate they get more efficient to run. A 5% variation in this parameter moves the 2040 total in our base case by 0.63%.

The spend figure is the cost of computing used during the year, as if rented, not the cost of buying the machines, so it should not be compared with reported capital expenditure, which is several times larger. Its starting level is built from what the laboratories disclose; only its growth rate comes from a published forecast.

3. Scenarios

We focus on three scenarios in our model, with varying assumptions on adoption, usage intensity, efficiency and training spend together, in combinations that hold as descriptions of a world rather than as arithmetic extremes. The span between Low and High is an envelope of plausible paths, not a confidence interval. Each scenario is described in Section 2, and the inputs behind them are set out in the key assumptions table below.

22. Jegham, N., Abdelatti, M., Elmoubarki, L. & Hendawi, A., "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference", arXiv:2505.09598 v2, May 2025.

4. Key assumptions

The inputs that most affect our estimate. Values for 2025 come from the sources shown in the basis column. Values for 2040 are our own judgement: informed by the evidence available today, but best guesses rather than precise estimates. That uncertainty is why we present three scenarios rather than a single figure.

Parameter	Unit	2025	Low 2040	Base 2040	High 2040	Basis
Inference						
Share of population using AI	%	16.3	55	70	90	Microsoft AI Economy Institute, H2 2025. ²³ Measured on the working-age population and applied to the total population, on the view that lower use among over-65s and higher use among under-15s broadly offset each other.
User mix: casual / moderate / heavy	%	70 / 25 / 5	48 / 36 / 16	40 / 35 / 25	32 / 38 / 30	Sized against industry estimates of the number of professional developers worldwide and global knowledge workers (ILO). ²⁴
Blended output tokens per user per day	'000 tokens	15	396	1,256	3,348	Derived from NBER research on ChatGPT usage, ²⁵ provider disclosures and Jegham et al. (2025)
Energy per 1,000 output tokens	Wh	3.8	0.36	0.18	0.08	Derived from Jegham et al. (2025), excludes data-centre overheads. Including overheads, a typical request today uses about 6 Wh, within a published range from 0.2 Wh for a median consumer chat prompt to over 33 Wh for a reasoning model on a long prompt. ²⁶
Energy per second of generated video	Wh	18.5	1.75	1.54	1.05	ML.ENERGY Leaderboard v3.0, median across five video models. ²⁷ Excludes data-centre overheads.
Hardware efficiency gain	%/yr	22	8	10	13	Based on Epoch AI analysis that machine-learning chip performance per watt improved about 28% a year over 2010–24. ²⁸
Software efficiency gain	%/yr	25	8	10	12	Our base case compounds to around a 3x gain by 2030, consistent with published estimates of near-term gains from model and serving improvements. ²⁹

23. Microsoft AI Economy Institute, "Global AI Adoption in 2025: A Widening Digital Divide", January 2026.

24. Developers: SlashData, "Global developer population trends 2025: how many developers are there?", April 2025. Knowledge workers: International Labour Organization, "World Employment and Social Outlook: May 2025 Update", which puts managers, professionals, and technicians and associate professionals at around a fifth of global employment.

25. Chatterji, A. et al., "How People Use ChatGPT", NBER Working Paper 34255, September 2025.

26. Elsworth et al., "Measuring the environmental impact of delivering AI at Google Scale", arXiv:2508.15734, August 2025.

27. ML.ENERGY, "ML.ENERGY Leaderboard v3.0", University of Michigan. See Chung, J.-W. et al., "The ML.ENERGY Benchmark: Toward Automated Inference Energy Measurement and Optimization", NeurIPS Datasets and Benchmarks, 2025.

28. Sevilla et al., "Can AI scaling continue through 2030?", Epoch AI, August 2024.

29. Oviedo, F. et al., "Energy use of AI inference, efficiency pathways, and test-time scaling", Joule, 2026.

Parameter	Unit	2025	Low 2040	Base 2040	High 2040	Basis
PUE	x	1.25	1.16	1.13	1.10	Our 2025 figure of 1.25 draws on two benchmarks: disclosed hyperscale fleet averages of about 1.08-1.17 ³⁰ and an all-facility average of 1.54. ³¹ Most generative AI runs in hyperscale facilities, so we set it closer to the hyperscale end. By 2040 the whole estate converges on today's hyperscale range, reaching the most efficient fleet averages in the high case.
Training						
Global training compute spend	2025 \$USb	31	46	95	195	Estimated from disclosed laboratory compute spend, scaled up to a global figure.
Spend growth: 2026-30 / 2031-35 / 2036-40	%/yr	-	10 / 3 / 2	20 / 8 / 3	25 / 17 / 5	Base 2026-30 from Bloomberg Intelligence's AI training infrastructure forecast. ³²

5. Principal limitations

- All three scenarios start from the same estimate of energy per token. That estimate is inferred rather than metered, and its source carries an uncertainty of roughly a factor of two. An error there would move every scenario together, so our Low-High range does not capture it.
- The model has no economic layer. It estimates how much electricity a given level of use requires, not whether that use will be paid for. If the value of generative AI disappointed, nothing inside the model would slow demand; that judgement enters only through the choice of scenario.
- Our boundary will become harder to observe. We count tokens generated, whatever product they sit behind. As generative AI becomes embedded in other software, it will be harder to separate from other data-centre demand in published figures, and so harder to check our estimates against them.

30. Company disclosures: Google 1.09 (2025), "Power usage effectiveness"; Amazon Web Services 1.14 (2025), "AWS Cloud"; Meta 1.08, as reported in "Meta data center electricity consumption hits 14,975GWh", DCD, July 2026; Microsoft 1.17 (FY2025), "2026 Environmental Sustainability Report".

31. Uptime Institute, "Global Data Center Survey 2025", July 2025.

32. Cerebras Systems, "Form S-1 Registration Statement", April 2026, citing Bloomberg Intelligence.

Important information

The opinions expressed are those of the author(s) as of the date indicated and may change based on market and other conditions. The accuracy of the content and its relevance to your client's particular circumstances is not guaranteed.

This market commentary has been prepared for general informational purposes by the team, who are part of Macquarie Asset Management (MAM), the asset management business of Macquarie Group (Macquarie), and is not a product of the Macquarie Research Department. This market commentary reflects the views of the team and statements in it may differ from the views of others in MAM or of other Macquarie divisions or groups, including Macquarie Research. This market commentary has not been prepared to comply with requirements designed to promote the independence of investment research and is accordingly not subject to any prohibition on dealing ahead of the dissemination of investment research.

Nothing in this market commentary shall be construed as a solicitation to buy or sell any security or other product, or to engage in or refrain from engaging in any transaction. Macquarie conducts a global full-service, integrated investment banking, asset management, and brokerage business. Macquarie may do, and seek to do, business with any of the companies covered in this market commentary. Macquarie has investment banking and other business relationships with a significant number of companies, which may include companies that are discussed in this commentary, and may have positions in financial instruments or other financial interests in the subject matter of this market commentary. As a result, investors should be aware that Macquarie may have a conflict of interest that could affect the objectivity of this market commentary. In preparing this market commentary, we did not take into account the investment objectives, financial situation or needs of any particular client. You should not make an investment decision on the basis of this market commentary. Before making an investment decision you need to consider, with or without the assistance of an adviser, whether the investment is appropriate in light of your particular investment needs, objectives and financial circumstances.

Macquarie salespeople, traders and other professionals may provide oral or written market commentary, analysis, trading strategies or research products to Macquarie's clients that reflect opinions which are different from or contrary to the opinions expressed in this market commentary. Macquarie's asset management business (including MAM), principal trading desks and investing businesses may make investment decisions that are inconsistent with the views expressed in this commentary. There are risks involved in investing. The price of securities and other financial products can and does fluctuate and an individual security or financial product may even become valueless. International investors are reminded of the additional risks inherent in international investments, such as currency fluctuations and international or local financial, market, economic, tax or regulatory conditions, which may adversely affect the value of the investment. This market commentary is based on information obtained from sources believed to be reliable, but we do not make any representation or warranty that it is accurate, complete or up to date. We accept no obligation to correct or update the information or opinions in this market commentary. Opinions, information, and data in this market commentary are as of the date indicated on the cover and subject to change without notice. No member of the Macquarie Group accepts any liability whatsoever for any direct, indirect, consequential or other loss arising from any use of this market commentary and/or further communication in relation to this market commentary. Some of the data in this market commentary may be sourced from information and materials published by government or industry bodies or agencies, however this market commentary is neither endorsed nor certified by any such bodies or agencies. This market commentary does not constitute legal, tax accounting or investment advice. Recipients should independently evaluate any specific investment in consultation with their legal, tax, accounting, and investment advisors. Past performance is not indicative of future results.

This market commentary may include forward looking statements, forecasts, estimates, projections, opinions and investment theses, which may be identified by the use of terminology such as "anticipate", "believe", "estimate", "expect", "intend", "may", "can", "plan", "will", "would", "should", "seek", "project", "continue", "target" and similar expressions. No representation is made or will be made that any forward-looking statements will be achieved or will prove to be correct or that any assumptions on which such statements may be based are reasonable. A number of factors could cause actual future results and operations to vary materially and adversely from the forward-looking statements. Qualitative statements regarding

political, regulatory, market and economic environments and opportunities are based on the team's opinion, belief and judgment.

Other than Macquarie Bank Limited ABN 46 008 583 542 ("Macquarie Bank"), any Macquarie Group entity noted in this document is not an authorised deposit-taking institution for the purposes of the Banking Act 1959 (Commonwealth of Australia). The obligations of these other Macquarie Group entities do not represent deposits or other liabilities of Macquarie Bank. Macquarie Bank does not guarantee or otherwise provide assurance in respect of the obligations of these other Macquarie Group entities. In addition, if this document relates to an investment, (a) the investor is subject to investment risk including possible delays in repayment and loss of income and principal invested and (b) none of Macquarie Bank or any other Macquarie Group entity guarantees any particular rate of return on or the performance of the investment, nor do they guarantee repayment of capital in respect of the investment.

Past performance does not guarantee future results.

Diversification may not protect against market risk.

Market risk is the risk that all or a majority of the securities in a certain market – like the stock market or bond market – will decline in value because of factors such as adverse political or economic conditions, future expectations, investor confidence, or heavy institutional selling.

International investments entail risks including fluctuation in currency values, differences in accounting principles, or economic or political instability. Investing in emerging markets can be riskier than investing in established foreign markets due to increased volatility, lower trading volume, and higher risk of market closures. In many emerging markets, there is substantially less publicly available information and the available information may be incomplete or misleading. Legal claims are generally more difficult to pursue.

Currency risk is the risk that fluctuations in exchange rates between the US dollar and foreign currencies and between various foreign currencies may cause the value of an investment to decline. The market for some (or all) currencies may from time to time have low trading volume and become illiquid, which may prevent an investment from effecting positions or from promptly liquidating unfavourable positions in such markets, thus subjecting the investment to substantial losses.

Infrastructure companies may be subject to a variety of factors that may adversely affect their business or operations, including high interest costs, high leverage, economic slowdowns, surplus capacity, increased competition, commodity prices, regulatory and political developments, difficulty raising capital, and terrorist acts or political actions, and general changes in market sentiment.

Compound annual growth rate (CAGR) is the constant annual rate at which a value would need to grow to move from its starting level to its ending level over a stated period, with each year's growth building on the last. It smooths year-to-year variation, so it does not show the path taken or the risk involved.

Terawatt-hour (TWh) is a unit of energy equal to one trillion watt-hours. Gigawatt (GW) is a unit of power equal to one billion watts. Average continuous draw is annual electricity use divided by the 8,760 hours in a year.

Behind-the-meter generation is electricity produced on the customer's side of the utility meter and used on site, rather than delivered through the grid.

Macquarie Group, its employees and officers may act in different, potentially conflicting, roles in providing the financial services referred to in this document. The Macquarie Group entities may from time to time act as trustee, administrator, registrar, custodian, investment manager or investment advisor, representative or otherwise for a product or may be otherwise involved in or with, other products and clients which have similar investment objectives to those of the products described herein. Due to the conflicting nature of these roles, the interests of Macquarie Group may from time to time be inconsistent with the interests of investors. Macquarie Group entities may receive remuneration as a result of acting in these roles. Macquarie Group has conflict of interest policies which aim to manage conflicts of interest.

All third-party marks cited are the property of their respective owners.

© 2026 Macquarie Group Limited

Pathways

For more information, or to speak to the author of this edition, Audrey Lee, please contact your Macquarie Asset Management Relationship Manager.