Policy Brief
Issue n°12/25 | March 2025

How DeepSeek has changed artificial intelligence and what it means for Europe

Bertin Martens

Bertin Martens (bertin.martens@bruegel.org) is a Senior Fellow at Bruegel

Executive Summary

BY MID-2024, ARTIFICIAL INTELLIGENCE large language models (LLMs) were running into diminishing returns to scale in training data and computational capacity. LLM training began to shift away from costly pre-training to cheaper fine-tuning and allowing LLMs to 'reason' for longer before replying to questions.

FINE-TUNING USES CHAIN-OF-THOUGHT (CoT) training data that includes questions and the logical steps to reach correct answers. This increases the efficiency of learning for smaller AI models, such as DeepSeek. CoT data can be extracted from large 'teacher' LLMs to train small 'student' models.

THESE CHANGES SHIFT the cost structure of AI models from high pre-training costs to lower fine-tuning costs for model developers and more inference costs for users. While smaller models are cheaper to use, a positive AI demand effect is likely to exceed the negative price effect. Price competition between models will increase, resulting in tighter margins for AI firms. Specialised models can still fetch premium prices.

CHEAPER LLMS ARE an opportunity for European Union companies to catch up in building smaller AI models and applications on top of LLMs. Increased demand for AI services will require more investment in computing infrastructure, including in the EU. Investing in large LLMs and the corresponding hyperscale infrastructure is riskier, especially as price competition between models increases.

KNOWLEDGE EXTRACTION BETWEEN AI models puts pressure on model developers to protect their investments against free-riding by others. It also creates a dilemma for policymakers: should they favour free-riding to promote competition and innovation, or should they clamp down and reinforce monopoly rents to stimulate investment in AI models? Past policy will not be an appropriate response in a world that offers vastly expanded opportunities for knowledge pooling and innovation at lower cost.

Recommended citation
Martens, B. (2025) 'How DeepSeek has changed artificial intelligence and what it means for Europe', *Policy Brief* 12/2025, Bruegel

1 Introduction: enter DeepSeek

The start of 2025 was marked by several major announcements related to artificial intelligence. The release of the DeepSeek (2025) AI model on 22 January blew a trillion-dollar hole in the stock market[1], on the basis that China's DeepSeek would substantially undercut American AI giants. DeepSeek was soon followed by many copy-cat small and cheap AI models. Markets concluded somewhat prematurely that DeepSeek broke the AI model scaling law and would undermine the rationale for heavy investment in AI computing infrastructure.

But can small AI models really perform as well as big models, without access to huge quantities of the expensive Nvidia AI processor chips that dominate the sector? Big-tech AI firms were not impressed by this market turbulence and doubled down on their AI infrastructure spending[2]. Just a week before the DeepSeek release, OpenAI and Oracle announced a $100 billion to $500 billion AI infrastructure investment – dubbed Stargate – to catch up with big-tech firms[3]. Two weeks later, the European Union announced its own €200 billion AI investment initiative[4].

DeepSeek is innovative, but in line with model evolution over the past half year – not an unexpected revolution

This Policy Brief aims to go beyond the DeepSeek hype. It analyses innovations in AI models over the past half year and examines the economic implications for AI companies and policymakers, in particular in the EU. It argues that DeepSeek is innovative, but in line with model evolution over the past half year – not an unexpected revolution. It still fits into the 'transformer' generative AI or large language model (LLM) paradigm of the last eight years (Vaswani *et al*, 2017).

Nevertheless, it has set in motion major changes in AI business models. The cost structure of AI models has shifted away from upstream pre-training costs towards more downstream fine-tuning costs for model developers and more computational 'reasoning' or inference costs to respond to the queries of end users. Moreover, AI models are increasingly free-riding on, and extracting knowledge from, each other. Price competition between AI models has further increased because smaller models are cheaper to operate. End-user costs per token[5] have dropped precipitously, making it more difficult for AI companies to extract revenue and make a profit from AI services. This creates a dilemma for AI developers. Should they protect their models against free-riding by others, if at all possible, or should they resolutely go for more innovation to stay ahead of competitors?

All these changes may create opportunities for EU innovators to catch up in the global AI race. EU policymakers should support smaller AI models, built on top of large models, to reap innovation and productivity gains from AI, without risky investment in large foundational models.

1 Sinéad Carew, Amanda Cooper and Ankur Banerjee, 'DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss', *Reuters*, 28 January 2025, https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/.

2 Mike Isaac, 'Meta to Increase Spending to $65 Billion This Year in A.I. Push', *New York Times*, 27 January 2025, https://www.nytimes.com/2025/01/24/technology/meta-data-center.html; Jordan Novet, "Microsoft reiterates plan to invest $80 billion in AI, but may "adjust our infrastructure in some areas"", *CNBC*, 24 February 2025, https://www.cnbc.com/2025/02/24/microsoft-reiterates-plan-to-invest-80-billion-in-ai-.html.

3 See OpenAI press release of 21 January 2025, 'Announcing the Stargate Project', https://openai.com/index/announcing-the-stargate-project/.

4 See European Commission press release of 11 February 2025, 'EU launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence', https://ec.europa.eu/commission/presscorner/detail/en/ip_25_467.

5 A token is a measure of the text length of AI training data and outputs. A token is equivalent to about three quarters of a word.

2 Beyond the DeepSeek hype: what has changed and why?

DeepSeek first caught the attention of AI developers towards the end of 2024 with its Version 3 'classic' LLM, which scored reasonably well on standard performance benchmarks for mathematics, computer coding and understanding of texts[6], though still below the most advanced models. This was quite an achievement for a Chinese AI model developer with limited access to the most advanced Nvidia AI computing chips[7]. Among several innovations, DeepSeek massively scaled up the use of existing 'mixture-of-experts' model training technology (Jacobs *et al*, 1991). Mixture of experts was invented in 1991 when the earliest AI models ran into computing capacity constraints with then state-of-the-art hardware. It divides a large model into sub-models, each with their own expertise. This saves on memory capacity and computing costs.

The follow-up DeepSeek R1 model subsequently ranked in the top five on the Chatbot Arena LLM Leaderboard[8], a subjective performance ranking of AI models by users. DeepSeek R1 is popular because it displays its reasoning steps explicitly when replying to a question. It also performs well on more objective benchmarks. To understand how DeepSeek R1 could achieve this, the rapidly evolving structure of LLMs over the last year must be examined.

From the start of the current generation of LLMs in 2017 until mid-2024, leading LLMs invested heavily in the AI model pre-training phase that consumes huge volumes of training data (or tokens) and computing power to estimate billions of internal parameters. LLMs exhibited a scaling law (Kaplan *et al*, 2020): better model performance requires more training data and more computing capacity. That drove investment in hyperscale computing facilities, with specialised AI processor chips. Nvidia's graphic processing units (GPUs) hold a near-monopoly in that market segment (Martens, 2024a).

By mid-2024, LLM developers realised they were close to exhausting the available stock of human-generated training data

By mid-2024, LLM developers realised they were close to exhausting the available stock of human-generated training data, or even future growth in that stock. That seems to put limits on expanding the scale of models and triggers decreasing returns from further expansion (Villalobos *et al*, 2024). A first solution to this was to let AI models produce their own 'synthetic' training data. But synthetic pre-training data soon ran into diminishing returns too because it was highly correlated with the original data and offered very little additional training information (Shumailov *et al*, 2024).

A better solution was to produce structured Q&A datasets that document the chain-of-thought (CoT) or 'reasoning' process that explains how to respond to a question to produce a correct answer. Models can then be fine-tuned in a second phase, after pre-training, by learning how to 'reason' their way to a correct answer. Previously, fine-tuning had been done on a smaller scale with human feedback. CoT datasets replaced human feedback and considerably shortened the learning cycle[9]. While CoT data can be curated by humans, it can be obtained much more cheaply and on a larger scale by an older technique called 'distillation' (Hinton *et al*, 2015), or extraction of knowledge from other LLM models. Distillation involves smaller 'student' models learning from larger and better 'teacher' models.

DeepSeek used its own LLM, DeepSeek V3, in combination with other LLMs including Ali-

6 For an overview of widely used tests and a comparison of model performance on these tests, see for example https://epoch.ai/data/ai-benchmarking-dashboard.

7 DeepSeek claims to have access to only 2,500 high-end Nvidia processor chips. However, Amodei (2025) estimated that it has access to mix of about 50,000 of more- and less-advanced Nvidia chips. This is still an order of magnitude less than the largest hyperscale AI infrastructure announced by Meta and XAI, but sounds more realistic for training an LLM like V3.

8 See 'Chatbot Arena LLM Leaderboard', https://lmarena.ai/?leaderboard.

9 Maarten Grootendorst, 'A Visual Guide to Reasoning LLMs: Exploring Test-Time Compute Techniques and DeepSeek-R1', 3 February 2025, https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms.

baba's Qwen, Meta's Llama and probably OpenAI's o1 reasoning model, to generate 800,000 CoT examples that it used for reinforcement learning in the fine-tuning stage of DeepSeek R1. Reinforcement learning requires less computational resources than pre-training. That allowed DeepSeek (2025) to claim that its model cost only a fraction of major LLMs – a claim that sent stock markets, already very nervous about over-valued AI stocks, tumbling (see section 1).

Since then, copy-cat models have emerged that apply the same procedure to all kinds of CoT training datasets, even more cheaply than DeepSeek R1. In February 2025, Muennighoff *et al* (2025) reached a new record, fine-tuning an AI model on just 1000 CoT examples, taking just seven hours to train. Prime Intellect (Mattern and Hagemann, 2025), an AI start-up, generalised this training procedure to generate CoT training data from a wide range of existing AI models. In fact, DeepSeek R1 showed openly what OpenAI had already been doing secretly for a while with its series of reasoning models, starting with OpenAI o1 in September 2024[10]. O1 was an expensive model when the combined cost of pre-training and fine-tuning is considered. DeepSeek's R1 low-cost claim was based on the cost of fine-tuning only.

The arrival of DeepSeek marks the start of a new cycle in AI model development. Models build on top of each other, with recursive development between models to accelerate AI learning at ever lower cost. The approach to model training based on training-data collection, pre-training and fine-tuning by a single firm is being replaced by a more horizontal networked model, potentially reshaping the AI field by proliferation and interaction – some call it free-riding – between AI models, combining distillation of knowledge and distributed training across models (Moussa *et al*, 2025). AI systems are being pulled out of a small number of big compute silos and sucked into a universe of smaller, more-powerful models with lower end-user inference costs. They are derived from and built on top of larger models with an additional fine-tuning training phase.

AI systems are being pulled out of a small number of big compute silos and sucked into a universe of smaller, more-powerful models with lower end-user inference costs

So far, LLMs have been *"stochastic parrots"*[11]: they require statistically representative samples of examples to learn from. OpenAI o1, DeepSeek and subsequent smaller reasoning models are gradually escaping from this statistical constraint by using more information-rich CoT training data that substantially increase the gradient of the learning curve and reduces computation requirements accordingly. Humans pursue the same learning strategy, with great success – humans are able to learn from just one or a few examples because of an ability to detect similarities between examples through lateral thinking (De Bono, 1970). Children learn to count from concrete examples – counting apples for example. Once they understand the basic principles at a more abstract level, they can apply the same counting rules in many other settings.

Chollet (2019) constructed the Abstraction and Reasoning Corpus (ARC), an AI benchmark that measures this type of thinking. Children score high on this benchmark but it is still difficult for AI models because they are still relatively weak in abstraction. OpenAI's o3 model is among the first to score well on the ARC test (Chollet, 2024). True one-shot learning from a single example, or even zero-shot learning based purely on reasoning (Pourpanah *et al*, 2023) would finally overcome the statistical constraints of LLMs and turn them from parrots into true reasoners.

10 See OpenAI press release of 12 September 2024, 'Introducing OpenAI o1-preview', https://openai.com/index/introducing-openai-o1-preview/.

11 A label invented by linguist Emily Bender. See Elizabeth Weil, 'You Are Not a Parrot', *New York Magazine*, 1 March 2023, https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html.

3 Reasoning models and the AI model scaling law

Amodei (2025) rejected the notion that DeepSeek showed that smaller reasoning models break the AI scaling law and that there is consequently no need to further increase computing capacity and training data to improve the performance of AI models. He argued that DeepSeek reduced training costs eightfold, mainly by successfully and more efficiently applying existing training techniques, such as mixture-of-experts (Jacobs *et al*, 1991), under pressure from US export controls that resulted in a scarcity of advanced Nvidia AI processor chips in China. However, DeepSeek's performance is still below that of the most advanced models by a factor of two, according to Amodei (2025). Overall, the net efficiency gain is thus about fourfold. That puts DeepSeek on the trend line of gradual improvement in the performance of LLMs – but it is not a revolution. DeepSeek R1 development costs only cover the fine-tuning stage and exclude the cost of pre-training the underlying LLMs from which CoT reasoning was distilled. When DeepSeek R1 and V3 are considered together as a joint product, the scaling laws are not broken.

CoT datasets contain more structured information than pre-training data. This is very similar to the way students learn from handbooks and exercises that explain how to approach problems and find correct solutions. The same CoT-reasoning approach can be implemented beyond fine-tuning in the next 'inference' stage when models reply to user questions. Allowing a model more time to 'think' about how to respond to a user question enables it to explore a variety of alternative internal reasoning processes and select those that produce better replies.

Smaller models require fewer computations to respond to a given question. That makes them cheaper for end users. However, allowing them more time for internal reasoning to produce better replies increases demand for computing resources. The negative price effect attracts more users and more applications for AI models. That quantity effect dominates and results in a net cost increase for model end users, though not for model developers. Developers thus do not pay a large chunk of the costs, in the form of run-time costs for users. As a result, investment in AI computing infrastructure is expected to continue to grow because of growing user demand for AI-based services.

Until the end of 2024, LLM performance was governed by the pre-training scaling law: more data and computing power increased performance. That put AI model costs on an unsustainable exponential cost growth trajectory (Martens, 2024b). Limitations on the volume of available data resulted in declining marginal returns to scale. The arrival of DeepSeek and OpenAI's reasoning models shifted training away from pre-training to fine-tuning or post-training. That gave LLMs a new lease of economic life. Post-training scaling laws provide a welcome technological and economic reprieve from the pre-training scaling law. If anything, the steepness of the model learning curve has increased, as suggested by Grootendorst (2025) and Busbridge *et al* (2025). Shifting to fine-tuning with CoT data does not break the scaling law. Rather, it starts a new fine-tuning scaling-law cycle.

4 Impacts on the economics of AI

The AI cost structure is shifting from very high fixed costs for pre-training, to lower fixed costs for fine-tuning and finally to reasoning costs for users at the inference stage

As a result of all these changes in model-training technology, the AI cost structure has shifted downstream, first from LLM pre-training to second-stage fine-tuning, and further down to 'reasoning' at the inference or test-time stage when the model responds to user questions. Data costs have also decreased, from collecting huge primary pre-training datasets to collecting much smaller CoT datasets with reasoning examples, extracted from other LLMs.

This entails a shift from very high fixed costs for LLM developers for pre-training, to lower fixed costs for developers for fine-tuning, and finally to more recurrent reasoning costs for users at the inference stage. Smaller models, such as DeepSeek, have lower user costs per unit (token) of input (questions) and output (replies) at the inference stage because they are cheaper to operate than large models. However, the volume of inference computing increases when models think longer over a reply and when lower prices stimulate users to use more models and ask more questions.

The shift from pre-training to fine-tuning raises questions about the optimal combination of the two training methods. Busbridge *et al* (2025) provided an empirical estimate of this relationship in the form of *"distillation scaling laws"* that depend on the relative size and performance of teacher and student models. For a given amount of training data, there is an optimal ratio between the size of student and teacher models. When the performance gap between the two becomes too large, an overly capable teacher model produces worse student models (Busbridge *et al*, 2025). In this way, a computational equilibrium emerges between teacher and student models, or between original learning in the pre-training phase and derived or distilled learning in the reinforcement stage. That indicates limits to free-riding or distillation, or a balance between acquired and own learning in LLMs – again very similar to human learning. Putting a university professor before a primary school class would create a very wide performance gap. Letting primary school pupils teach each other would narrow the performance gap too far.

This increases the complexity of pricing of LLM use. Bergemann *et al* (2025) showed how AI model developers need to design pricing packages for users that take into account the volume and cost of input and output tokens during inference, as well as the duration of inference computations. Pricing can be linear or based on volumes and the value of computations and users' willingness to pay for that value.

Moreover, pricing needs to amortise the fixed costs of pre-training and reinforcement learning. If LLM developers operate in a monopolistic market in which model quality differences are important, they can mark-up prices above marginal inference costs to amortise the fixed costs of training. If they are in a competitive market with minor quality differences, mark-up pricing becomes difficult. Developers may also underprice inference to increase their market shares. The rapid evolution and turnover of LLM rankings on the Chatbot Arena Leaderboard indicates vigorous quality competition between LLMs. Models may become more differentiated as they upload specialised CoT training data in the second fine-tuning phase. However, any monopolistic training-data advantage will be quickly eroded in an AI ecosystem in which models increasingly interact with, and extract knowledge from, each other.

On one hand, marginal cost pricing at the user inference stage, combined with fixed costs for pre-training and fine-tuning or reinforcement learning, undermines the economic sustainability of AI business models. The only way out for LLM developers is to reduce price competition, produce higher value-added services that can be sold at premium prices or bundle their LLMs with other services for which they have a stronger market position. For example, Microsoft bundles LLMs with office productivity software. Meta bundles LLMs with targeted advertising services on its social media platforms. Apple bundles LLMs with its devices. Anthropic is developing specialised AI business services[12].

12 Cristina Criddle, 'Anthropic set to focus on business users in search for new revenues', *Financial Times*, 18 March 2025, https://www.ft.com/content/97e0ab06-8d83-4918-9079-3ed935bc1c63.

On the other hand, monopolistic business strategies draw the attention of antitrust and competition policymakers in the EU and elsewhere. EU policymakers have several instruments at their disposal to keep monopolistic behaviour at bay. Should large AI companies be designated as 'gatekeepers' – or dominant, hard-to-avoid platforms – under the EU Digital Markets Act (DMA, Regulation (EU) 2022/1925) and bundle LLMs with core platform services under that act, the obligations of DMA Art 6§7 apply[13]. This forces gatekeepers to ensure free third-party access and interoperability with all software and hardware components separately, as an unbundled or vertically separable service. That would undo any attempts by AI developers to make the market more monopolistic.

Would this be a good strategy for competition policymakers? The risk of this approach is that it would undermine the economic sustainability of AI business models, at least for big-tech AI firms that meet the DMA's quantitative threshold criteria. Smaller firms that use AI would escape from this and be able to price their specialised AI services at a mark-up above marginal user inference costs. That would be a perfectly pro-competitive policy, unless smaller AI firms with 'student' AI models continue to depend on larger 'teacher' LLMs. Eroding the economic viability of the latter would also affect the performance of the former. This is a policy area in the making that requires careful observation of market developments.

Another important question in this newly emerging AI ecosystem is how tolerant AI companies will be of mutual distillation and free-riding. Intellectual property rights do not apply to LLMs. The terms of use of most LLMs prohibit use for designing competing products. This may be hard to enforce, unless technical protection measures can be designed that make distillation difficult.

Schrepel and Potts (2024) underscored the commons nature of much AI technology. Software used for AI models is shared via open-source platforms such as GitHub and Hugging Face, though open-source user licenses come with a variety of conditions. AI modelling expertise is widely shared online through research papers and blogs. Model pre-training data is compiled from webpages and other text sources, though some of this data may violate copyright.

Beyond these more traditional commons resources for AI models, DeepSeek has laid bare a new commons dimension: a network of interacting AI models. That creates a decentralised pool that combines overlapping knowledge sets from all these models. No single party controls this pool but everyone has access to it, unless legal or technical protection measures would prevent this. Allowing access to that pool opens up vast opportunities to accelerate innovation and more knowledge accumulation. It would further accelerate competition between AI companies and between countries in a tense geopolitical context.

This creates a policy dilemma. Open access and hyper-pooling of knowledge in all AI models would be highly beneficial from an overall societal-welfare perspective. It would accelerate innovation and stimulate competition and minimise user inference costs. However, open access and free-riding could undermine investment incentives in LLM pre-training and fine-tuning. Open-access networked models may also ultimately undermine the quality of model responses as original learning is discouraged in favour of copying (Rogers, 1988).

For economists, this is a well-known dilemma: striking a balance between monopolistic rent-seeking behaviour and dynamic innovation. The former is an unavoidable evil in order to produce the benefits of the latter. The same policy dilemma is already playing out in pre-training of LLMs that require access to huge volumes of copyright-protected training data, giving rise to tensions between rightsholders and AI developers (Martens, 2024c).

Policymakers will have to find a new balance between these opposites. Because AI reduces the cost of innovation, the degree of protection of monopolistic investment rents can be lowered, compared to the pre-AI world. But it cannot be eliminated entirely. Defensive technical-protection measures could be installed that track the type and frequency of user questions, or that profile users, to detect distillation methods. These could be combined with

13 AI models are not yet classified as a core platform service under the DMA. But this may change.

user-price discrimination, with prices increased for longer data extractions (Bergemann *et al*, 2025).

In addition to these defensive measures, model developers can change strategies and build more-exclusive vertically integrated and specialised services on top of their AI models – and then charge premium prices for these services. Whether these technical and commercial strategies are sufficient to create viable AI business models remains to be seen.

5 How do changing AI technologies affect AI development in the EU?

Unlike the US, the EU does not have home-grown big-tech companies that operate hyperscale computing infrastructure and generate sufficient revenues from their global business models to cover the high costs of LLM development. OpenAI and Oracle's $100 billion to $500 billion Stargate AI infrastructure investment announcement[14] signalled that these second-tier players want to become big-tech AI players with access to hyperscale computing infrastructure. It also signals that OpenAI is trying to wean itself off its $13 billion 'coopetition' agreement with Microsoft, exchanging access to Microsoft computing infrastructure in return for use of OpenAI's ChatGPT model in Microsoft services (Martens, 2024a). Whether OpenAI will succeed depends on its ability to rapidly scale up its $5 billion (2024) annual revenues, to catch up with Microsoft's revenues of $245 billion.

In contrast with Stargate's private financing, a European Union initiative announced in January 2024, AI Factories[15], was all about injecting several billions of taxpayers' money into existing government-owned scientific supercomputers. Fortunately, in time for the Paris AI Summit in February 2025[16], the EU realised that it needed to be more ambitious and to pull in private investors. French President Emmanuel Macron announced €109 billion in mostly private investment in state-of-the-art computing infrastructure, some of it financed by the same sovereign wealth funds as the Stargate initiative[17]. In parallel, European Commission President Ursula von der Leyen announced a €200 billion investment in AI infrastructure[18], also a mixture of private and public money. How all these quickly-compiled big-figure investment announcements will roll out in practice remains to be seen of course. The announcements lack details on commitments.

DeepSeek has given smaller AI companies the prospect of competing with big-tech AI models at a much lower AI-model cost, avoiding heavy investment in hyperscale infrastructure. However, as argued above, that is only part of the story. DeepSeek depends on knowledge and training data extraction from larger models. It is important here to distinguish between the scale of an AI model and the scale of computing infrastructure.

Large-scale LLMs will still be needed, if only as teacher models for smaller student AI models. But fine-tuning and reasoning at the inference stage can be accomplished by smaller models. That creates an opportunity for smaller AI firms to enter the market.

The scale of AI computing infrastructure will still have to expand as cheaper and smaller AI models spread and the positive quantity effect outpaces the negative price effect. More computing power will be needed for inferencing. Since user inferencing infrastructure is best

14 See footnote 3.

15 See European Commission, 'AI Factories', undated, https://digital-strategy.ec.europa.eu/en/policies/ai-factories.

16 See https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia.

17 Jacob Wulff Wold, 'France unveils €109 billion AI investment plan', *Euractiv*, 10 February 2025, https://www.euractiv.com/section/tech/news/france-unveils-e109-billion-ai-investment-plan/.

18 See footnote 4.

located geographically close to users, there are certainly good opportunities to invest in that type of infrastructure in the EU. To what extent that will happen in large concentrated computing centres or in a more distributed way across smaller facilities, possibly with edge computing in consumer and firm devices, remains to be seen. The Commission's announcement in February 2025 appears to be a mixture of larger and smaller AI infrastructure projects.

Whether the EU can successfully build its own foundational LLMs is highly uncertain. Economically successful LLMs require not only AI infrastructure and skilled staff. They also require business models that can earn a rate of return on the huge fixed investment costs. Big-tech firms can bundle LLMs with their existing global services business models and earn additional revenue. For smaller EU firms, that is much more difficult when they operate in a market with strong price competition and have to sell their AI services at marginal cost. They may be better off below the technology frontier, building specialised niche market AI applications on top of existing LLMs (Martens, 2024a).

Developing a layer of smaller and more specialised fine-tuned AI models and applications on top of existing LLMs would be sufficient to spur innovation and productivity growth in the EU. Whether smaller AI firms can charge a profitable marked-up price to users to pay their due share in LLM development costs will depend on competition in their niche markets. As discussed above, the shift in the AI model cost structure makes it even more difficult to amortise the fixed costs of LLM pre-training in a highly competitive end-user market in which user inferencing services are sold at marginal cost.

6 Policy conclusions

The release of the DeepSeek R1 model brought into the open a shift in AI model technology that had already started in mid-2024, when larger AI models ran into an economically unsustainable trajectory with exploding costs and decreasing returns to scale. AI model costs shifted away from pre-training to fine-tuning, with the help of reinforcement learning on more logically structured CoT data, and to longer 'reasoning' time at the inference stage when models respond to user questions. This also marks a shift from high fixed costs for model developers to more recurrent costs for users during inference.

At the same time, competition between models remains very vigorous. User inference is often priced at marginal computation cost, leaving little or no profit margins for model developers, unless they can bundle their models with specialised services. Moreover, increased knowledge extraction or 'distillation' between large LLM 'teacher' models and smaller 'student' models results in a degree of free-riding that may erode the business models of LLM developers.

Policymakers should be careful to avoid putting their faith in pre-AI policy recipes, and should be open to taking a pro-innovation stance

Policymakers face a dilemma. Should they protect static monopolistic rents for model developers as an incentive to invest in LLM development and avoid free-riding by smaller models? Or should they favour smaller derived AI models and model applications that free-ride and build on top of existing LLMs, as a way of promoting more innovation? This is a difficult balancing exercise in a rapidly evolving technological setting. Policymakers should be careful to avoid putting their faith in pre-AI policy recipes, and should be open to taking a pro-innovation stance.

While DeepSeek cast doubts over the wisdom of investment in AI computing infrastructure, we have argued in favour of continuing to do so because demand for AI services is very likely to continue to expand and require ever more computing capacity. The EU has announced several AI investment initiatives, with a mixture of private and taxpayer money and a range of smaller and larger infrastructure clusters. A focus on smaller infrastructure and AI models and applications would certainly contribute to lifting productivity growth in the EU. Large-scale infrastructure for foundational LLMs is far riskier. Along with experienced

staff and adequate computational resources, it requires large-scale business models to earn a return on these huge fixed investment costs. Building applications on top of existing LLMs may be less risky, especially because of the shift in AI cost structures and increasing competition squeezes prices and profit margins.

References

Amodei, D. (2025) 'On DeepSeek and Export Controls', blog post, January, available at https://darioamodei.com/on-deepseek-and-export-controls

Bergemann, D., A. Bonatti and A. Smolin (2025) 'The Economics of Large Language Models: Token Allocation, Fine-Tuning, and Optimal Pricing', mimeo, available at https://arxiv.org/abs/2502.07736

Busbridge, D., A. Shidani, F. Weers, J. Ramapuram, E. Littwin and R. Webb (2025) 'Distillation Scaling Laws', mimeo, available at https://arxiv.org/abs/2502.08606

Chollet, F. (2019) 'On the measure of intelligence', mimeo, available at https://arxiv.org/abs/1911.01547

Chollet, F. (2024) 'OpenAI o3 Breakthrough High Score on ARC-AGI-Pub', *ARC Prize*, 20 December, available at https://arcprize.org/blog/oai-o3-pub-breakthrough

De Bono, E. (1970) *Lateral thinking*, Penguin

DeepSeek (2025) 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning', mimeo, available at https://arxiv.org/abs/2501.12948

Hinton, G., O. Vinyals and J. Dean (2015) 'Distilling the Knowledge in a Neural Network', mimeo, available at https://arxiv.org/abs/1503.02531

Jacobs, A., M. Jordan, S. Nowlan and G. Hinton (1991) 'Adaptive Mixtures of Local Experts', *Neural Computation* 3(1): 79-87, available at https://doi.org/10.1162/neco.1991.3.1.79

Kaplan, J., S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child ... D. Amodei (2020) 'Scaling Laws for Neural Language Models', mimeo, available at https://arxiv.org/abs/2001.08361

Martens, B. (2024a) 'Why artificial intelligence is creating fundamental challenges for competition policy', *Policy Brief* 16/2024, Bruegel, available at https://www.bruegel.org/sites/default/files/2024-07/PB%2016%202024.pdf

Martens, B. (2024b) 'Catch-up with the US or prosper below the tech frontier? An EU artificial intelligence strategy', *Policy Brief* 25/2024, Bruegel, available at https://www.bruegel.org/sites/default/files/2024-10/PB%2025%202024.pdf

Martens, B. (2024c) 'Economic arguments in favour of reducing copyright protection for generative AI inputs and outputs', *Working Paper* 09/2024, Bruegel, available at https://www.bruegel.org/system/files/2024-04/WP%2009%20040424%20Copyright%20final_0.pdf

Mattern, J. and J. Hagemann (2025) 'Decentralized Training in the Inference-Time-Compute Paradigm', *Prime Intellect*, 21 January, available at https://www.primeintellect.ai/blog/intellect-math

Moussa, H., A. Akhavain, S. Hosseini and B. McCormick (2025) 'Distributed Learning and Inference Systems: A Networking Perspective', mimeo, available at https://arxiv.org/abs/2501.05323

Muennighoff, N., Z. Yang, W. Shi, L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès and T. Hashimoto (2025) 's1: Simple test-time scaling', mimeo, available at https://arxiv.org/abs/2501.19393

Pourpanah, F., M. Abdar, Y. Luo, X. Zhou, R. Wang, C. Peng Lim, X. Zhao Wang and J. Wu (2023) 'A Review of Generalized Zero-Shot Learning Methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(3), available at https://doi.org/10.1109/TPAMI.2022.3191696

Rogers, A. (1988) 'Does biology constrain culture?' *American Anthropologist* 90(4): 819-831, available at https://www.jstor.org/stable/680759

Schrepel, T. and J. Potts (2024) 'Measuring the openness of AI foundation models: competition and policy implications', *Information & Communications Technology Law*, available at https://doi.org/10.1080/13600834.2025.2461953

Shumailov, I., Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson and Y. Gal (2024) 'AI models collapse when trained on recursively generated data', *Nature* 632: 755-759, available at https://www.nature.com/articles/s41586-024-07566-y

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin (2017) 'Attention Is All You Need', mimeo, available at https://arxiv.org/abs/1706.03762

Villalobos, P., A. Ho, J. Sevilla, T. Besiroglu, L. Heim and M. Hobbhahn (2024) 'Will we run out of data? Limits of LLM scaling based on human-generated data', *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235: 49523-49544, available at https://raw.githubusercontent.com/mlresearch/v235/main/assets/villalobos24a/villalobos24a.pdf