Equity Insights

AI Democratised: What lower cost of compute means for the investment landscape

June 2025

For professional investors only

HSBC Asset Management

Table of contents

This commentary provides a high-level overview of the recent economic environment and is for information purposes only. It is a marketing communication and does not constitute investment advice or a recommendation to any reader of this content to buy or sell investments nor should it be regarded as investment research. It has not been prepared in accordance with legal requirements designed to promote the independence of investment research and is not subject to any prohibition on dealing ahead of its dissemination. The views expressed above were held at the time of preparation and are subject to change without notice. Any forecast, projection or target where provided is indicative only and not guaranteed in any way. HSBC Asset Management accepts no liability for any failure to meet such forecast, projection or target.

Foreword

> The AI revolution is no longer just about creating the smartest machine, but about scaling access and delivering value to the most hands at the lowest cost.

Pierin Menzli
Global CIO, Equity
HSBC Asset Management

Welcome to the May 2025 edition of our Equity Insights publication. In the context of a quickly evolving AI landscape that has altered investor considerations around prospective long-term 'winners', we present our latest research and implications around opportunities in the global technology sector.

Our Equity insights publication of last October discussed our thinking around the growing importance of inference and expectations for model improvements to reduce the cost of compute and result in a broadening out of the equity landscape, positioning the software sector to potentially lead the next stage of the investment cycle. Subsequently, advancements in AI large-language models from China sent shockwaves through technology markets early in the year, shifting investor sentiment, particularly against the semiconductor and hardware companies that had led gains previously.

While the market reaction and corresponding volatility points to a lack of clarity on what these developments truly mean going forward, certainly part of this change in sentiment was the premise that cheaper, more efficient AI models will reduce the need for expensive hardware underneath it. We think this point requires much nuance.

Of course, much more market turmoil followed in response to trade policy announcements. The broad decline in US stocks has now been followed by a rapid rebound that has seen the tech sector as a whole regain its strength, leading us to now look at where global technology markets go from here.

We continue to believe that as in past platform shifts, software providers stand to accrue a large share of the benefits of this new technology by facilitating the delivery of it to enterprises and consumers, with adoption set to grow as costs decline.

We also think that declining costs in conjunction with more innovation and accessibility present particularly interesting considerations for emerging Asia, where technology providers have inherent advantages in scaling and commercialising AI with consumers. This extends beyond China tech stocks, which have been rewarded this year, to beneficiaries elsewhere in Asia also.

In this publication we discuss these thoughts in depth, and trust it offers useful insights as you consider implications for your investment strategies.

In a nutshell

- Reasoning models like DeepSeek R1 mark a shift from traditional predictive models to more advanced ones capable of analysing information, drawing conclusions and solving complex problems.

- An evolving AI landscape means beneficiaries are becoming less concentrated and the US no longer holds unchallenged dominance.

- While the US remains a leader in foundational AI development, China's engineering optimisations, scale advantages, and commercialisation potential position it as a significant player in the next phase of AI growth.

- Most importantly for the next phase of growth, falling costs of AI compute are democratising access to the technology. This trend is expected to unlock higher-value applications and broaden enterprise AI adoption globally.

- All of this supports our expectation for software firms to lead the next phase of the AI investment cycle, driving the integration of generative AI into products and solutions.

- With very broad technology ecosystems and superapps, emerging Asia presents opportunities for scaling and commercialisation of AI with consumers that are not replicable in developed markets.

- Ultimately, the democratisation of AI presents much opportunity for this new technology to fuel earnings growth in emerging Asia and broader emerging markets. This could potentially erode the significant edge in profitability that fuelled historic outperformance by US tech firms over the last decade.

Evolution of the AI story

> A new AI model originating from China early in the year dramatically lowered the cost of AI compute, placing Chinese companies firmly into the AI value chain and sowing confusion among investors around what comes next.

Ed Conroy
Global Head of Equity Research

Nick Dowell
Senior Portfolio Manager,
Thematic Equities

While much of the volatility in US stock markets this year has been at the hands of trade policy developments, technology stocks initially came under pressure since the late January emergence of new AI models from China. Markets struggled to evaluate the implications, which at once dramatically lowered the cost of compute. The DeepSeek models – comparable to US models but requiring significantly less investment – contain several important breakthroughs around training, inference and how to improve efficiency by optimising the hardware used.

Given the focus on how cheap it had been to train the model, the hundreds of billions of dollars spent on infrastructure and training by the industry was suddenly cast in a negative light, putting ex-China IT stocks under selling pressure. Of course, significant outperformance from parts of the tech sector that fuelled the post-ChatGPT rally came with expectations of continued heavy investment in AI infrastructure. Accordingly, this part of the market was vulnerable to a wobble in confidence. We believe, though, that there are important misconceptions as to the true significance of the model developments. As demonstrated below, the arrival of this new innovation has not dampened interest in the use of existing models.
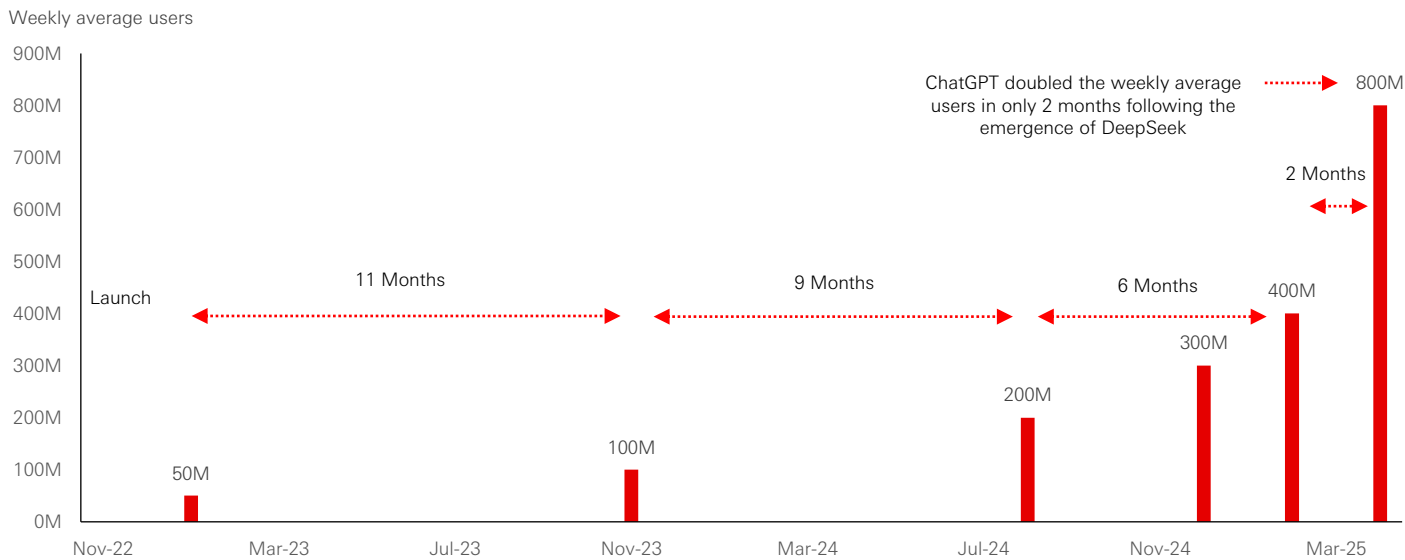
Figure 1: ChatGPT's growth has continued to accelerate post-DeepSeek

Weekly average users

Source: Barclays Research, company disclosures, HSBC AM as of April 2025.

DeepSeek's breakthroughs

While not the only one, DeepSeek's R1 model represents a shift from traditional predictive LLMs that provide answers based on pre-existing data, to reasoning models that can analyse information, draw conclusions, and generate solutions to complex problems. This capability is akin to human reasoning, where individuals assess situations, consider various factors and arrive at informed decisions. For example, when posed with a question a traditional model might provide an immediate answer based on its training data. In contrast, a reasoning model breaks the question down into components, evaluates the information and arrives at a more accurate conclusion. This iterative process enhances the quality of the answers generated, making reasoning models particularly valuable in complex scenarios.

The ability to reason is essential for AI systems to operate effectively in real-world situations. In enterprise settings, reasoning models can streamline processes such as onboarding of new employees for instance, facilitating an AI agent capable of interacting with various systems to action relevant steps (i.e. issue access pass, procure relevant laptop type, etc.). Of course, the potential goes beyond this. By automating tasks and making informed decisions based on available data, these models reduce friction and enhance operational efficiency.

Yet, DeepSeek's real ingenuity lies in engineering optimisations that reduce AI compute costs while maintaining performance. This has been made possible through architectural innovations such as Mixture of Experts (MoE), Multi-Head Latent Attention (MLA), and unique communication protocols enabling high efficiency even on lower performance chips like the Nvidia H800 that was permissible for export to China.

In terms of impact going forward, DeepSeek has open-sourced its work, democratising access to cutting-edge reasoning models and empowering developers globally to iterate, adapt, and deploy AI at scale. This will be helpful in enhancing the technology and addressing issues such as AI 'hallucinations'. Greater accessibility should also translate to more businesses leveraging the value of these models, supporting more demand-driven investments, fuelling further innovation and growth that could be instrumental in reducing inference costs.

OpenAI has also launched reasoning models but had essentially kept the process closed. This changed post-DeepSeek, with the subsequent promise of an open-source model to be released in the coming months – signifying a competitive shift that supports accessibility and creates beneficiaries down the AI supply chain.

The level of optimisation carried out by DeepSeek was not necessary for model builders in the US, who had greater access to compute.

Figure 2: Resources for training of models

	GPT-3 (Open AI)	GPT-4 (Open AI)	Llama-3.1 (Meta)	DeepSeek V2	DeepSeek V3
Model parameters (bn)	175	1,800	430	236	670
Training tokens (bn)	3,700?	13,000	15,000	8,100	14,800
Training chipsets	V100	A100	H100	H800	H800
No. of GPUs	1,024	25,000	16,000	2,000+	2048
Training cost (US$mn)	4.6mn	>100mn	60mn	5mn+ ??	5.57mn

Source: HSBC AM, company data. April 2025

The cost of AI: misconceptions and realignment

The launch of DeepSeek's R1 garnered significant attention due to its widely-circulated figure of $5.5 million in training costs. For many reasons we think this underestimates the full investment and misses the point. The lasting impact is the release of important breakthroughs to the model community that could substantially lower the cost of inference and thus accelerate AI adoption.

One improvement employed by DeepSeek is how the algorithms work and how data is used. Using specific, very targeted chunks of data to generate answers vastly improves efficiency, speeding up cost reduction in these models.

Today, inference costs are declining by 10x annually on average. Reflecting the classic Jevons Paradox (i.e. as compute becomes cheaper and more efficient, consumption increases), a rise in consumption going forward should be due to:

- Increased user base (hundreds of millions using LLMs daily),

- More time spent per session (more complex reasoning),

- More compute per request.

> Jevons Paradox states that as we use a resource more efficiently, consumption of that resource goes UP, not down.

Lowering the cost of inference — the process of generating reasoning from AI models — speeds all of this. While a reasoning model will solve a complex task more accurately, it can require 150x more compute than a basic LLM, per an example shared by Nvidia's CEO Jensen Huang at their recent GPU Technology Conference. This is as much as 400 times that of a standard Google search.

The trajectory of hardware development should also support the path to cost reduction. Nvidia's new Blackwell and upcoming Rubens GPUs are expected to significantly reduce the total cost of ownership for inference operations. As these hardware advancements come to market over the next two years, reasoning models will become more accessible and scalable, unlocking higher-value applications and broadening enterprise AI adoption.

This is a net positive for the entire AI value chain, including hyperscalers, infrastructure providers and emerging software leaders who are positioned to benefit from the broader deployment of generative AI models.

In the short term, lower cost of compute should benefit the software sector relative to the semiconductor sector. As the software sector invests and integrates Generative AI into their products, inference is an operating cost to them – anything that can reduce this will be welcomed. This shift has started to be reflected in equity returns, with a rotation away from massive semiconductor outperformance generated in the initial post-ChatGPT rally.

In the context of previous platform shifts such as the mobile-internet cycle, while the semiconductor industry has been the primary benefactor from a flood of investment to date, the impact on the software sector is still in its early stages. Our view remains that substantial lowering of the cost of compute is a necessary condition for a move to the next stage of the investment cycle to take place.

For the hyperscalers and infrastructure providers, the impact becomes nuanced. Some have continued with their platform approach, allowing models to reside on their infrastructure, thus monetising usage and inference. Others have developed their own models as well as their own compute for internal purposes. Rapid advancements and commoditisation of compute is more likely to be a headwind for the latter approach and a tailwind for the former.

China joins the AI investment story

Up to now, the US has dominated the AI revolution, benefiting from deep capital markets, first-mover advantages, and robust ecosystems. Massive profitability of US tech firms has also supported investment in the technology. As of 2024, Nasdaq profits were four times the global average. However, that gap is now narrowing and brings regional broadening into consideration as part of the next stage of the investment cycle.

Today, US tech capex-to-sales ratios are above levels seen in the 2000 dot-com bubble. If those massive capital investments don't yield proportionate revenue growth, margins may compress and valuations reset. This could further this year's reallocation of global capital towards underrepresented regions which aligns with our investment outlook of markets and economies 'spinning around'.

China may not have led in foundational AI innovation, but their engineering prowess, scale advantages through superapps, and potentially easier path to commercialisation position them as major beneficiaries of the next wave of AI development. Amidst the rally in China technology stocks post-DeepSeek, this rerating still leaves China large cap tech at roughly half the price of US large cap tech relative to earnings. Accordingly, if monetisation and growth is realised, there is ample room for further rerating.

Although US companies are still leading globally in terms of original innovations, we believe China could lead globally in terms of engineering optimisations, production and commercialisation at scale. This creates a large growth runway for China's internet/cloud platforms, where consumer-facing LLM-powered applications can lead to improving monetisation and cost.

> Significant benefits to the Chinese economy and company fundamentals are still to come, but it appears to be a matter of time.

Figure 3: We are in early stages of growth in China cloud spend (Cloud spend as % of GDP)
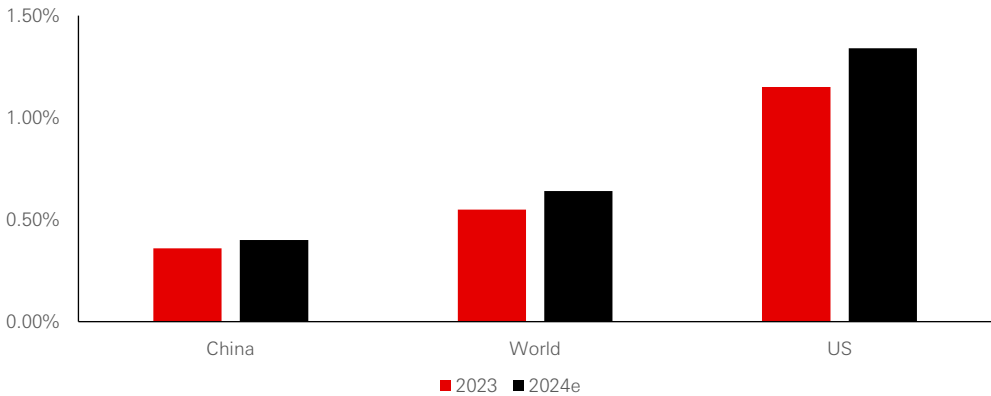
Source: Statista, Gartner, HSBC Research estimates. March 2025.

Demand for AI-driven solutions in China has surged this year since the release of DeepSeek. In conjunction, the use of open-source AI is paving the way for faster application development and innovation, presenting clear opportunities for independent AI application developers.

As early beneficiaries, the leading cloud providers are integrating DeepSeek R1-like models into their cloud infrastructure, creating new value propositions while enhancing operational efficiencies. Meanwhile, growing AI adoption creates beneficiaries down the local supply chain, with local hardware manufacturers and software developers filling gaps created by US export restrictions.

Figure 4: China's tech investment supports local advances

(100 million yuan)

Source: National Bureau of Statistics of China, as of 24 January 2025.

Figure 5: China semiconductor self-sufficiency

Source: Gartner, WSTS, Morgan Stanley Research estimates, February 2025

As AI use cases grow, the demand for customised and locally manufactured compute components is expected to rise. Furthermore, proliferation of AI-enabled applications on consumer devices is likely to shorten upgrade cycles, creating a favourable environment for hardware providers and component suppliers alike.

Ultimately, the value of AI model developments lies in how businesses use AI to improve their existing products and services over time. We can look to Meta as an example of how AI has already been monetised within a large consumer platform. Leveraging AI, Meta has automated much of its advertising process in the past two years. Advertisers can now use AI to generate ad content and develop advertising strategies that outperform human efforts. Further benefits include an AI-powered matching system that ensures a better advertising targeting rate.

Figure 6: Meta's advertising revenue and y-o-y growth (USDm)
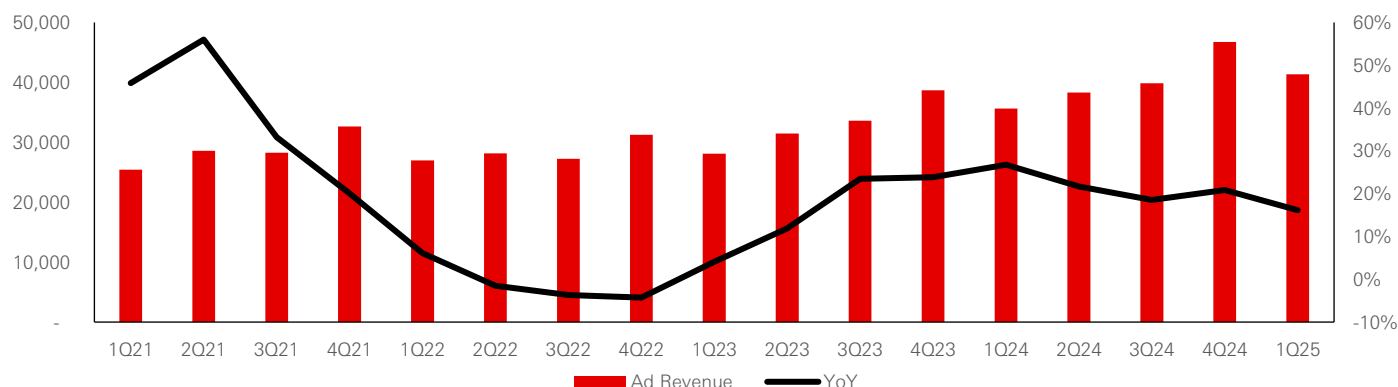
Source: HSBC Research, Company data, HSBC AM, March 2025.

With the help of DeepSeek, Chinese internet platforms are looking to implement similar improvements to their business models, with potential for ever larger impact thanks to their massive ecosystems. Over time, this could evolve into full-fledged AI agents capable of delivering advanced solutions across different domains. Vertical internet players such as online healthcare, education and travel platforms are also leveraging AI for personalized consulting services, creating new value proposition for consumers.

Beyond internet companies, a broad swathe of industries are becoming fast movers in integrating large language models into their operations. Likewise, government institutions are integrating AI, supporting expected growth in private AI cloud projects for state owned enterprises and government agencies.

Beneficiaries of growing AI adoption extend further into emerging Asia

Superapps in Asia can seamlessly facilitate the monetisation of AI functionality with consumers, presenting an opportunity that is not replicable in developed markets. As AI is integrated across these large platforms, enhancements should include more targeted advertising to a very broad user base, better e-commerce experiences for these consumers, and more efficient generation of entertainment content including games (allowing greater production of differentiated content for consumption). This supports the promise of AI leverage in Asia picking up steam and supporting earnings growth as the cost of compute continues to reduce.

Separately, as lower cost of compute drives greater enterprise adoption of generative AI globally, India becomes an interesting beneficiary for its own reasons. Commonly dubbed as 'the world's back office', a large part of India's technology sector consists of IT services. As system integrators, they stand to benefit from large scale adoption of Gen AI across enterprises.

Currently, global enterprise adoption is still transitioning from proof of concepts to larger scale adoptions as use cases evolve. Consulting services providers incorporating AI solutions, like Accenture, TCS and Infosys, have spoken about Gen AI gaining scale in their orderbooks. However, it is not at the core of offerings, still primarily focused on enhancing skills, capabilities and reskilling of employees – it largely remains on the periphery of solutions rather than a core driver.

At the same time as spending on AI infrastructure – data processing and storage – has progressed at a rapid pace, there has been a slowdown in spend on Indian IT services over the last year. This points to spend on AI buildout consuming budgets, leaving smaller slices of the pie for IT Services. Even the hyperscalers, with their near limitless budgets, have cut IT services spends as they have grown their substantial AI infrastructure capex. Figure 4 reflects this disparity in spend for stand-alone application development and maintenance services versus infrastructure spend globally.

> India could become an interesting beneficiary of increased enterprise Gen AI adoption.

Figure 7: 2024 Growth in ADM services
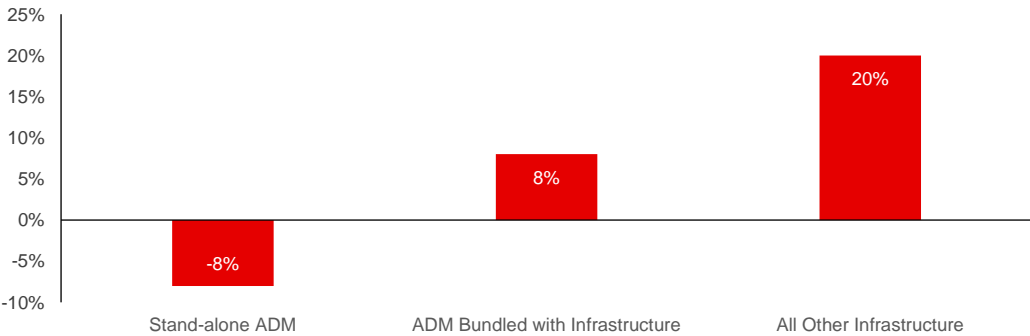
Source: Information Services Group Index, January 2025.

While DeepSeek may not be adopted in the west, the ideas around optimisation and doing more with less will certainly permeate US tech. This means the cost of adoption will continue to come down and more domain specific LLMs will likely be developed. Here, Indian IT can participate in optimising LLMs, even with lower processing power. Provided the US macro environment supports continued corporate spending ahead, we expect a revival of discretionary spending for Indian IT Services as the Gen AI story evolves.

What comes next

DeepSeek's rise has muddled the AI investment narrative and marked a turning point in certain ways. Yet, such developments in the technology are to be expected, and per Jevons Paradox, only contribute to the ultimate conclusion of AI's reach expanding across companies and consumers worldwide.

With reasoning models now in the hands of billions, the AI race is no longer just about who builds the smartest machine, but who gets it into the most hands at the lowest cost. And while the US remains the leader in foundational AI development, its unchallenged dominance is fading.

In the long arc of technological revolutions—from steam to silicon to intelligence— the winners are those who adapt, scale, and distribute. We see several implications for equity portfolios. In terms of market segments, we continue to believe that software firms will lead the next stage of the investment cycle. It is this segment that must facilitate the delivery and use of AI both in enterprises and with consumers, as declining compute costs expand the opportunity set for software solutions.

Nonetheless, much investment is still needed in the hardware to achieve the desired model capabilities and efficiencies. Accordingly, a long growth runway for high-performance GPU chips that have powered the AI revolution remain, along with the less powerful but more flexible ASIC chips whose functionality may become more useful as model efficiencies increase and use cases expand.

Regionally, we anticipate that emerging Asia will gain greater investor attention. With the launch of the DeepSeek models broadening and accelerating access to leading AI capabilities, adoption in these markets is supported by inherent advantages in monetising the technology. And as the cost of compute declines, operational gains and enhanced product and service offerings should feed through to earnings growth in these markets. Accordingly, we could see the significant edge in profitability that fuelled historic outperformance by US tech firms over the last decade begin to erode.

> We expect a rotation in 'winners' of the AI trade.

Figure 8: Net profit growth over a decade

Rebased to 100, Jan 2015

Figure 9: Consensus EPS growth (IBES) - Nasdaq from 4x to 2x world growth in 2025e

%

Source: Refinitiv, Datastream, MSCI, HSBC AM, March 2025.

Important information

- in Germany by HSBC Global Asset Management (Deutschland) GmbH which is regulated by BaFin (German clients) respective by the Austrian Financial Market Supervision FMA (Austrian clients);

- in Hong Kong by HSBC Global Asset Management (Hong Kong) Limited, which is regulated by the Securities and Futures Commission. This video/content has not be reviewed by the Securities and Futures Commission;

- in India by HSBC Asset Management (India) Pvt Ltd. which is regulated by the Securities and Exchange Board of India;

- in Italy and Spain by HSBC Global Asset Management (France), a Portfolio Management Company authorised by the French regulatory authority AMF (no. GP99026) and through the Italian and Spanish branches of HSBC Global Asset Management (France), regulated respectively by Banca d'Italia and Commissione Nazionale per le Società e la Borsa (Consob) in Italy, and the Comisión Nacional del Mercado de Valores (CNMV) in Spain;

- in Malta by HSBC Global Asset Management (Malta) Limited which is regulated and licensed to conduct Investment Services by the Malta Financial Services Authority under the Investment Services Act;

- in Mexico by HSBC Global Asset Management (Mexico), SA de CV, Sociedad Operadora de Fondos de Inversión, Grupo Financiero HSBC which is regulated by Comisión Nacional Bancaria y de Valores;

- in the United Arab Emirates, Qatar, Bahrain & Kuwait by HSBC Global Asset Management MENA, a unit within HSBC Bank Middle East Limited, U.A.E Branch, PO Box 66 Dubai, UAE, regulated by the Central Bank of the U.A.E. and the Securities and Commodities Authority in the UAE under SCA license number 602004 for the purpose of this promotion and lead regulated by the Dubai Financial Services Authority. HSBC Bank Middle East Limited is a member of the HSBC Group and HSBC Global Asset Management MENA are marketing the relevant product only in a sub-distributing capacity on a principal-to-principal basis. HSBC Global Asset Management MENA may not be licensed under the laws of the recipient's country of residence and therefore may not be subject to supervision of the local regulator in the recipient's country of residence. One of more of the products and services of the manufacturer may not have been approved by or registered with the local regulator and the assets may be booked outside of the recipient's country of residence.

- in Peru: HSBC Bank USA NA has an authorized representative by the Superintendencia de Banca y Seguros in Perú whereby its activities conform to the General Legal Financial System - Law No. 26702. Funds have not been registered before the Superintendencia del Mercado de Valores (SMV) and are being placed by means of a private offer. SMV has not reviewed the information provided to the investor. This document is for the exclusive use of institutional investors in Perú and is not for public distribution;

- in Singapore by HSBC Global Asset Management (Singapore) Limited, which is regulated by the Monetary Authority of Singapore. The content in the document/video has not been reviewed by the Monetary Authority of Singapore;

- in Switzerland by HSBC Global Asset Management (Switzerland) AG. This document is intended for professional investor use only. For opting in and opting out according to FinSA, please refer to our website; if you wish to change your client categorization, please inform us. HSBC Global Asset Management (Switzerland) AG having its registered office at Gartenstrasse 26, PO Box, CH-8002 Zurich has a licence as an asset manager of collective investment schemes and as a representative of foreign collective investment schemes. Disputes regarding legal claims between the Client and HSBC Global Asset Management (Switzerland) AG can be settled by an ombudsman in mediation proceedings. HSBC Global Asset Management (Switzerland) AG is affiliated to the ombudsman FINOS having its registered address at Talstrasse 20, 8001 Zurich. There are general risks associated with financial instruments, please refer to the Swiss Banking Association ("SBA") Brochure "Risks Involved in Trading in Financial Instruments;

- in Taiwan by HSBC Global Asset Management (Taiwan) Limited which is regulated by the Financial Supervisory Commission R.O.C. (Taiwan);

- in Turkiye by HSBC Asset Management A.S. Turkiye (AMTU) which is regulated by Capital Markets Board of Turkiye. Any information here is not intended to distribute in any jurisdiction where AMTU does not have a right to. Any views here should not be perceived as investment advice, product/service offer and/or promise of income. Information given here might not be suitable for all investors and investors should be giving their own independent decisions. The investment information, comments and advice given herein are not part of investment advice activity. Investment advice services are provided by authorized institutions to persons and entities privately by considering their risk and return preferences, whereas the comments and advice included herein are of a general nature. Therefore, they may not fit your financial situation and risk and return preferences. For this reason, making an investment decision only by relying on the information given herein may not give rise to results that fit your expectations.

- in the UK by HSBC Global Asset Management (UK) Limited, which is authorised and regulated by the Financial Conduct Authority;

- and in the US by HSBC Global Asset Management (USA) Inc. which is an investment adviser registered with the US Securities and Exchange Commission.

- in Uruguay, operations by HSBC's headquarters or other offices of this bank located abroad are not subject to Uruguayan inspections or regulations and are not covered by warranty of the Uruguayan state. Further information may be obtained about the state guarantee to deposits at your bank or on www.bcu.gub.uy.