HOW CAN AI POLICY SUPPORT ECONOMIC GROWTH?

A report prepared for Microsoft

15 APRIL 2024

WWW.FRONTIER-ECONOMICS.COM

Contents

frontier economics

frontier economics

EXECUTIVE SUMMARY

Objectives and scope of this report

New Artificial Intelligence models capable of generating images, videos and text ("generative AI") have pushed the boundaries of AI capabilities, quickly gaining significant adoption, and showing the potential to generate substantial economic benefits. These advances have taken place through a complex value chain that includes the following "layers":[1]

- **Compute layer:** building and providing access to high-performance computing including the supply of specialist chips and datacentre infrastructure.
- **Foundation layer:** using the above computing resources along with vast quantities of data and specialist software tools to train "foundational" large language models; and
- **Application layer:** building and distributing software applications that use foundation models, such as chat-based assistants, co-pilots or plug-ins. Building these applications often involves "fine-tuning" a foundation model to specific tasks that could potentially be in almost any sector of the economy including healthcare, telecommunications, finance, retail, professional services, visual media, utilities, and more.

Given the substantial resources and capabilities required to be at the forefront of this market, it is likely that public policy will play an important role in determining to what extent the UK will be active in and benefit from generative AI. However, policy resources are finite and therefore it is crucial that they are invested in a way that maximises their return for taxpayers.

Therefore, the key question for policymakers is: how should government prioritise alternative strategic options to support the growth of generative AI in the UK? That is to say, how can policy best unlock value across the generative AI value chain and should government support be targeted towards specific layers of the value chain?

This report aims to provide policymakers with a framework to answer these questions, and an initial application of the framework based on existing evidence.

Overall assessment

The initial application of our economic framework suggests that Government should prioritise polices that aim to make the UK a leader in the generative AI application layer. This represents the greatest near-term opportunity for the UK, where a range of policy actions can help remove barriers and leverage the UK's existing capabilities, including:

- direct funding and access to finance for start-ups and scale-ups developing generative AI applications;

[1] Please note that the division of the value chain into three layers is a simplification and each layer includes distinct products, services and capabilities, as described in section 2 of this report.

- initiatives to promote AI safety in key sectors where the UK is well-placed to develop generative AI applications, given existing skills and export capabilities. These sectors are likely to include financial services, health, biotech and professional services;
- supporting the skills ecosystem in key areas for application development, such as data engineering and prompt engineering; and
- programmes to facilitate collaboration between generative AI start-ups and potential data providers.

Compute capacity is a key dependency for all other layers, however the time frames for increasing national capacity are longer-term. It is therefore important to consider how best to ensure short-term capacity to unlock opportunities at other layers. This could include focussing on removing barriers to developing domestic compute capacity for uses where this is required: these include developing foundation models in areas with highly sensitive data (e.g. health), and the deployment of generative AI applications to end users (where latency must be minimised and therefore domestic compute capacity is crucial). where domestic data residency is a requirement. A longer-term strategy for the UK's high-performance computing would help ensure that future demand of compute for these and other users is met through private and public provision. It is unlikely that the UK will be in a position to become an exporter of compute capacity in the short to medium term.

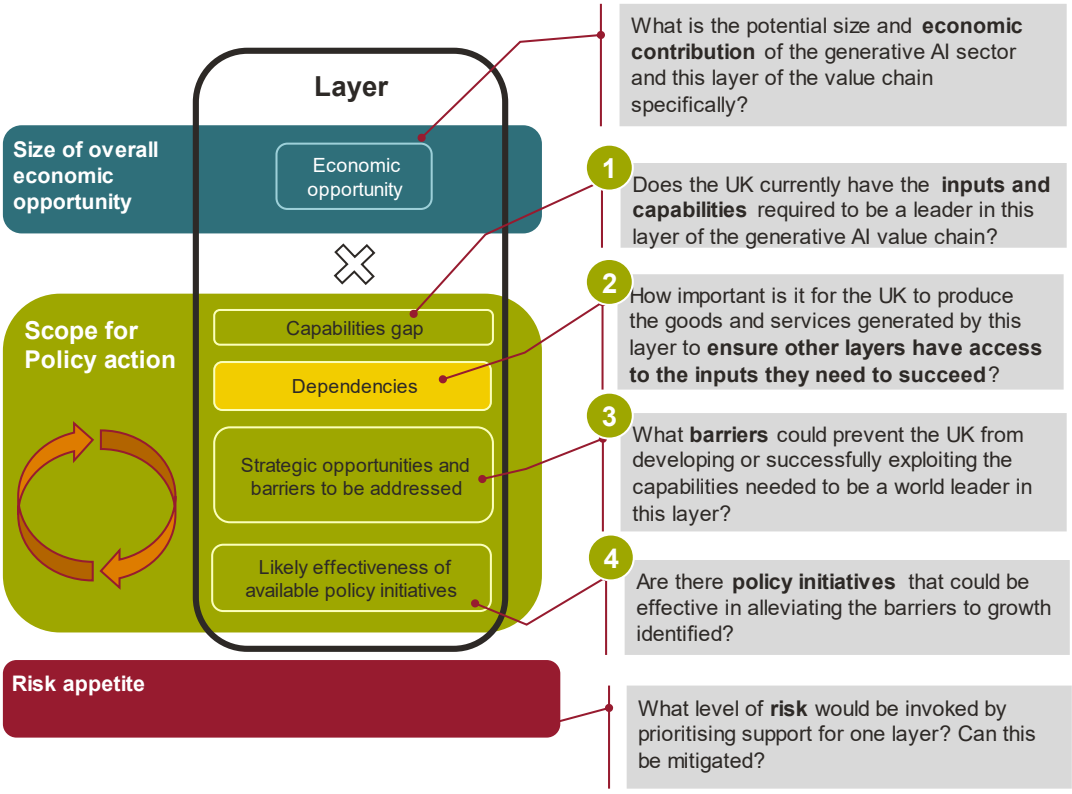
There are medium-term opportunities for government to support the development of the foundation layer. However, the impact on the wider ecosystem may be relatively small as there is a is already a wide range of foundation models available through varying levels of access.[2] Additionally, these investments would be higher risk, particularly as foundation model development involves high costs (relative to the development of gen AI applications) and is heavily dependent on highly specialised talent that is in short supply globally.

These conclusions assign higher priority to policy actions that aim to build on the UK's existing relative strengths and where there is greater evidence of barriers that could be effectively addressed by policies in the shorter term. An alternative approach that is also consistent with our framework would be to focus policy efforts on addressing areas where the UK's capabilities are currently weaker, and the UK is further away from being a leading international player. While such an alternative approach could potentially yield large benefits, the impact would be more uncertain and would only materialise in the longer term.

Our framework

Our initial assessment is based on a framework that involves analysing **four factors** for each layer of the generative AI value chain, and defining the level of risk involved in supporting that layer. The four factors are described in the figure below.

[2] CMA (2023). AI Foundation Models: Initial Report.



In principle, one might want to start by assessing the likely size of economic opportunity for the UK at each layer of the stack. However, generative AI technology and business models are still developing so it is very challenging to estimate precise figures for each layer. Therefore, we believe at this stage it is best to target policies to the layer where they are most likely to have a large impact – rather than attempting to target policies to the layer that has the highest value.

The assessment of the capabilities gap (factor 1) and dependencies (factor 2) helps to understand for each layer, the extent to which UK policy towards a layer might need to focus on:

- developing new capabilities because the UK's existing capabilities are currently limited or because the layer is a key input to other layers where the UK is better positioned; and/or
- maintaining and effectively leveraging existing capabilities because existing capabilities are strong but may need further support or targeting towards generative AI.

The next steps in the framework involves considering what barriers currently hinder development of each layer in the UK (factor 3) and the likely effectiveness of available policies to address those barriers (factor 4). This helps think through the specific rationales for policy actions, and rule out cases where government intervention may be less likely to have the desired impact (for example, if available policy options are not likely to be effective), or where it may have an impact that would not be truly additional (for example, if policy options are not addressing genuine barriers to generative AI development and simply displacing private sector investment).

In summary, the framework can be used to:

- Assess whether future policy actions should be targeted more heavily towards some layers compared to others;
- Assess what types of policy actions are most likely to be effective based on available evidence; and
- Identify key gaps in the evidence base that could be filled in the future to inform AI policy decisions

The policy options identified as higher priority could then be further assessed by estimating their expected benefits and costs (beyond the scope of this study).

The boxes below describe the key questions we have answered with current evidence to apply the framework, and the following section describes the key questions that we would recommend addressing in the future, as more evidence becomes available.

Capabilities

What do direct measures of relevant UK capabilities tell us about the UK's current capabilities in each layer of the generative AI value chain? Relevant areas include:

- The availability of relevant AI skills;
- The inclination of the UK's innovation and research ecosystem towards AI;
- Private sector willingness and ability to invest in AI related innovation; and
- The availability of VC investment for AI.

What do indirect measures of relevant UK capabilities, including evidence of comparative advantage in related sectors, tell us?

Dependencies

What does the architecture of the generative AI value chain and the current structure of this market tell us about current and likely future dependencies between the layers of the generative AI value chain?

To what extent is access to domestic compute capacity or UK trained foundation models required?

What factors (such as open source models) could help or hinder fulfilment of these requirements?

Barriers

What factors could prevent the private sector from fully realising the UK capabilities in each layer without policy support? Relevant factors could include:

- Knowledge spillovers from R&D not internalised by the private sector;
- Asymmetric information and capital market imperfections;
- Coordination problems;
- Underinvestment in skills; and
- Safety and security issues.

What barriers could arise from existing policy issues? For example:

- Planning restrictions and other constraints on data centres;
- Constraints to demand for generative AI in the public sector;
- Clarity of regulation; and
- Policies affecting access to data.

Policies

What are the key policy options that could address the barriers to growth identified?

How directly do these options address the barriers identified?

How quickly could they be expected to generate an impact?

Have they been shown to be effective in the past?

What wider benefits could they have, for example, what is their potential to have a broader impact on R&D spillovers and AI adoption?

Key questions for the future application of this framework

Evidence on the UK's capabilities for AI development, the relative importance of different barriers to growth faced by AI businesses, and the effectiveness of different policies is evolving rapidly. As such, the framework developed in this report is designed to be reapplied in the future as things evolve. Below we set out the key questions that we were not able to fully address in our initial application of the framework (either due to limited available evidence or the nascent and developing nature of the sector). Future applications of this framework could seek to address these questions through primary evidence gathering, such as stakeholder consultation, surveys or pilot studies.

- What are the most important inputs and capabilities for each layer of the generative AI value chain? For example, what specific skills are most important for each layer currently, and what skills are likely to be most important in the future?
- What is the current and likely future demand for and supply of UK based computing capacity (rather than global cloud computing)? To what extent may a lack of domestic compute capacity prove to be a bottleneck in the future?
- What will be the future role of open-source foundation models in both AI research and commercial application development? Will open-source models continue to provide efficient access to foundation models, including for fine-tuning to specific applications?

- Which are the most important barriers to the development of generative AI in the UK? For example, is access to compute a more significant barrier than access to finance? How does this differ by layer of the generative AI value chain?
- What is the economic impact and value for money of policies that aim to i) promote the safety of AI products, ii) develop the UK's digital infrastructure and iii) promote secure access to data?

Figure 1 Summary of results from framework application

Source: Frontier Economics

1 Introduction

Recent advances in AI have generated substantial policy maker interest in this fast-developing sector. This includes the development of new AI models capable of generating images, videos and text ("generative AI"). These advances have been delivered through a generative AI value chain that includes three broad layers:

- **Compute layer:** building and providing access to high-performance computing including the supply of specialist chips and datacentre infrastructure.
- **Foundation layer:** using the above computing resources along with vast quantities of data and specialist software tools to train "foundational" large language models; and
- **Application layer:** building and distributing software applications that use foundation models, such as chat-based assistants, co-pilots or plug-ins. Building these applications often involves "fine-tuning" a foundation model to specific tasks that could potentially be in almost any sector of the economy including healthcare, telecommunications, finance, retail, professional services, visual media, utilities, and more.

It is clear that generative AI has the potential to bring about significant innovation and economic growth, transforming many sectors of the economy, whilst also raising a number of challenges, including those related to safety, security, privacy, and intellectual property. Policy makers within governments are interested in how they can support the development and adoption of new AI technology in a safe and welfare enhancing manner.

However, generative AI continues to evolve at a rapid pace within a complex and interdependent value chain. There is considerable uncertainty about what capabilities and factors will be important for the development of the sector in both the short and longer term. As a result, it can be very challenging for government policy makers to understand whether and how they should take action to maximise the benefits of generative AI to the UK.

In this context, Frontier Economics has been commissioned by Microsoft to produce an economic analysis that can support policymakers in the UK to make better informed decisions about AI policy, with a focus on generative AI. The question this report aims to help answer is: what types of policies could be most effective in supporting the growth of the generative AI in the UK? Specifically, given the economics of the generative AI value chain, should government support be targeted towards specific layers of the stack?

In answering these questions, it is important to recognise that this sector is still relatively nascent and rapidly developing. It is therefore crucial to have a consistent and systematic framework that can both provide initial answers to these questions and also be used to update these answers over time as the generative AI market develops and more evidence becomes available.

This is the approach we follow in this report. Specifically, we have developed a framework that involves assessing four factors for each layer of the generative AI value chain. The four factors

we have identified are: **Capabilities**, **Dependencies**, **Barriers**, and **Policies**. This framework is summarised below.

It is worth noting that the development and adoption of generative AI could generate economic benefits through:

- Increasing productivity in existing sectors (including both efficiency gains and improving existing products)
- Generating new products and services
- Innovation spillovers onto other sectors and activities

Existing estimates place the global opportunity for generative AI at around $1 trillion to $3 trillion. This implies large potential value generation. However, technology and business models are still developing so it is very challenging to estimate precise figures for the market as a whole and even more challenging to value each layer. Therefore, we believe at this stage it is best to target policies to the layer where they are most likely to have a large economic impact. This is why we do not include the potential size of the economic opportunity for the UK at each layer as a factor to be assessed in our framework.

Our framework for assessing generative AI policy options

1. Capabilities

Key questions:

Does the UK currently have the inputs and capabilities required for businesses to compete effectively in each layer of the generative AI value chain? Given its capabilities (relative to other countries), does the UK have or could it feasibly develop a world leading comparative advantage in any of these layers?

Approach:

We assess direct measures of UK capabilities, including:

- Skills (both specialised skills and skills for adoption);
- Innovation and research outputs;
- Willingness and ability to invest in innovation; and
- Availability and ability to attract VC investment.

We also assess indirect measures of relevant UK capabilities, including evidence of comparative advantage in related sectors.

2. Dependencies

Key questions:

What are the interdependencies at each layer of the generative AI stack? Is there a case for supporting a particular layer in order to ensure sufficiency or security of supply of necessary inputs for other layers?

Approach:

We map out the current architecture of the generative AI supply chain and market structure and assess the current and likely future state of interdependencies between layers through literature review and consultation with industry stakeholders. This includes consideration of compute requirements in both the foundation and application layers and the role of open access foundation models.

3. Barriers

Key questions:

What are the barriers to growth in each layer of the generative AI value chain? That is to say, what are the reasons why the UK may not be able to fully reap the potential benefits from the sector (or mitigate potential risks) without government support?

Approach:

We consider theoretical reasons for and empirical evidence of barriers that may arise from misaligned incentives or informational constraints, including:

- Social benefits, such as R&D knowledge spillovers, not fully reflected in market incentives;
- Asymmetric and imperfect information affecting access to finance;
- Coordination challenges between different parts of the value chain;
- Underinvestment in skills; and
- Safety and security issues.

We also consider barriers that may arise from policy directly, including**:**

- Planning restrictions and other constraints on data centres;
- Constraints to demand for generative AI products in public sector organisations;
- Clarity of regulation; and
- Policies affecting access to data.

4. Policies

Key question:

What policy initiatives could be most effective in alleviating the barriers to growth identified?

Approach:

We identify key policy options available within the following areas:

- Supporting investment in science, R&D and broader innovative activities;
- Supporting the development of AI skills;
- Supporting access to finance for new ventures in the AI space;
- Promoting the safety of AI products and services and supporting justified trust in AI systems;

- Ensuring that the UK's digital infrastructure is AI-ready and future-proof;
- Promoting secure access to data.

For the key policy options in these areas we then assess:

- How directly they address the barriers identified;
- How quickly they could be expected to generate an impact;
- Whether they have been shown to be effective in the past;
- What benefits would they have beyond immediate effect on supply in the targeted layer, for example, what is their potential to have a broader impact on R&D spillovers and adoption?

Within this framework, the first two factors (Capabilities and Dependencies), identify whether there is a case for targeting support toward a particular layer of the generative AI stack either to:

1. Foster the development of that layer of the value chain as an area in which the UK can develop a world leading comparative advantage – meaning that the UK is a leading exporter of the technologies, products and services developed in that layer; or
2. Ensure that companies, researchers and government organisations operating in other parts of the value chain have access to the inputs they require.

The final two factors (Barriers and Policies), identify whether and how government policies could support the different layers of the generative AI stack. This framework is primarily based on the idea that the UK is best placed playing to its areas of comparative strength: it seeks to identify these areas of relative strength and identify the policies and inputs needed to support growth in these areas. However, the questions posed in our framework can also be used to assess what might be needed to develop a future comparative advantage in any area of the generative AI value chain.

While we do estimate the potential economic benefits that the generative AI sector could generate in the UK overall, we do not explicitly consider differences between layers of the generative AI stack in terms of gross value added as a factor within our framework. There is currently limited evidence to suggest that productivity differs significantly between the layers of the generative AI value chain and any such differences are likely to play a more minor role than the factors included in our framework for policy design.

In this report we offer an initial application of this framework to the UK based on the current state of the generative AI sector and available evidence. In doing so we demonstrate how this framework can be applied in practice and generate tentative conclusions for how UK policy could be best directed to support development of the generative AI sector. We consider each of the five questions in turn separately for each layer of the generative AI value chain. This analysis was conducted between August and November 2023 and may not reflect more recent market evolutions.

This study focusses primarily on innovation and industrial policy, broadly defined to include: government investment in R&D, grants and subsidies for private sector innovation, public procurement of digital technology, and investment in computing infrastructure. While there has been substantial recent discussion around AI safety and the regulation of AI, we do not analyse in detail possible options for the regulation of generative AI or AI safety. However, we do discuss the crucial role that AI safety regulation can play in supporting the development of the sector and the relative importance of this across different layers of the generative AI stack.

It should also be noted that this report focuses primarily on how policy can support economic activity in the generative AI sector – that is to say, the development of generative AI technologies, products and applications, through the compute, foundation and application layers of the generative AI value chain. We do not consider policies that could support adoption of these technologies by businesses and the general public. This is an important separate question but is beyond the scope of this report.

The remainder of this report is set out as follows:

- **Chapter 2 (Market Background)** describes the generative AI value chain and its component layers, as well as summarising the current market structure across these layers. We also provide some preliminary estimates of the potential economic impact of the generative AI sector.
- **Chapter 3 (Capabilities)** describes current capabilities in the UK that could be leveraged for growing the generative AI sector.
- **Chapter 4 (Dependencies)** discusses interdependencies between the different layers of the generative AI value chain.
- **Chapter 5 (Barriers)** analyses barriers to growth and potential reasons why the generative AI sector may not be able to realise its full potential benefits without government support.
- **Chapter 6 (Policies)** discusses potential policy options to support the development of the generative AI sector.
- **Chapter 7** draws overall implications and provides conclusions.

2 Defining the generative AI value chain

In this section we describe the generative AI value chain. We provide a relatively brief description, in order to inform the application of our framework to assessing AI policies from chapter 3 onwards. More extensive analysis of the value chain can be found in existing reports including the CMA's initial review of foundation models[3] or Bruegel[4]. In section 2.6, we also provide a brief description of the potential economic benefits linked to the growth of the generative AI sector.

We define the generative AI sector as including all individuals and organisations involved in "**generative AI production**". By "generative AI production", we mean all activities within the generative AI value chain, from the design of processing units to the monetisation of applications that rely substantially on a foundation model. We consider generative AI production as distinct from **generative AI adoption**, that is, the use of generative AI applications by consumers and businesses in the UK.

The distinction between generative AI production and adoption is not always clear-cut. For example, if a producer of mobile games uses generative AI, does this count as generative AI production or adoption? For the purposes of this report, we define the mobile games producer's activity as "AI production" if its applications rely "substantially" on a foundation model: in other words, if the applications would not be recognisable as essentially the same product without access to a foundation model. For example, if a crucial part of gameplay involves players generating their own characters or worlds using generative AI, the game could be considered as a "generative AI application" and the games producer could be considered as a "generative AI producer". On the opposite end of the spectrum, if the games producer uses generative AI coding assistants to support the work of its games developers, we would consider the games producer as an AI adopter, while the organisation providing the coding assistant as a generative AI producer. This definition could be improved and may need to be refined in the future as the capabilities of generative AI models and their use evolve. However, the application of our policy framework does not depend on where exactly one draws the line between production and adoption of generative AI.

The generative AI value chain is highly complex with many interconnected layers and sub layers. However, for simplicity we can think of it in terms of three main layers: (i) Compute layer; (ii) Foundation layer; and (iii) Application layer.

2.1 Compute layer

The **compute layer** refers to the computational inputs necessary for being able to process and generate data. The main activities are:

3 CMA (2023). AI Foundation Models: Initial Report

4 Bruegel (2023). Competition in generative artificial intelligence foundation models

- Designing, manufacturing and assembling hardware components (i.e. chips);
- Combining the hardware components into a data centre or a "supercomputer"; and
- Where relevant, making the computing infrastructure available through the internet as a cloud computing service.

Key inputs underpinning the compute layer are connectivity and electricity supply. In particular, access to renewable energy provision is particularly important for the largest data centre providers, which have all made significant commitments to the sustainability of their data centres.

Foundation model and AI application developers typically outsource their compute requirements to cloud/datacentre providers due to the significant fixed costs associated with datacentre infrastructure. There are multiple players in this market including Amazon Web Services, Microsoft Azure, Google Cloud, Oracle, IBM Cloud and others. Some developers and researchers may alternatively utilise individual "supercomputers" operated by universities or other institutions for their compute needs.

In terms of hardware, the scale of mathematical operations involved in developing and deploying generative AI models typically requires specialised processing units, especially at the training and pre-training stage. Large "grids" of high-end accelerator chips, such as Graphical Processing Units (GPUs) are used for this purpose, rather than general-purpose chips (Central Processing Units, CPUs). Due to the significant economies of scale involved and expertise required, a relatively small number of businesses have a significant presence in the manufacture of these chips. These businesses include NVIDIA, Intel and AMD. However, partly as a result of ongoing shortages of GPUs, a number of firms are investing in the development of alternative chips for AI workloads, such as Google's Tensor Processing Units, Amazon's Trainium and Inferentia chips (which will be used by Anthropic in training its future foundation models, and Microsoft's AI chips (Azure Cobalt, Azure Maia and Azure Boost).[5]

In this report, we are interested in the compute layer of the generative AI value chain in two ways:

- First, as a sector where the UK could play an active role. Playing an active role would mean that businesses that provide hardware, software and services required for the development and deployment models have a significant proportion of their operations in the UK. These businesses would operate in the UK not only for the purposes of serving UK customers but also to develop and deploy products and services provided to international customers in the global generative AI supply chain.

- Secondly, as an input into the foundational and application layer. The key issue here is whether producers of foundation models and applications operating in the UK would have

[5] See https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html; and https://www.aboutamazon.com/news/company-news/amazon-aws-anthropic-ai.

access to sufficient compute infrastructure for their operations. This compute infrastructure would not necessarily need to be provided by businesses that have a significant proportion of their operations in the UK. It could be provided by international businesses with limited presence in the UK and it could also include infrastructure provided by public sector organisations.

2.2 Foundation layer

The **foundation layer** involves the pre-training and training of foundational large language models using vast quantities of data and compute capacity, typically through a type of deep learning model called a transformer.

There are many different players active in the foundation layer. OpenAI's GPT model released in 2018 was the first publicly available foundation model. Since then, the CMA estimates that there have been around 160 foundation models developed and released.[6] Key players include OpenAI, Google, Meta, Anthropic, Microsoft, Adept, Stability AI and Hugging Face.

Much of the data for the training of foundation models comes from very large publicly available datasets, such as Common Crawl, The Pile, Project Gutenberg Corpus, LAION-400M, LAION5B, ROOTS, Red Pajama, RefinedWeb, and Starcoder. Some prominent foundation models such as LLaMA (Meta), GPT-3 (Open AI) and Stable Diffusion (Stability AI) were trained entirely on these sources.

Training of foundation models also requires software tools, libraries and resources that facilitate various stages of the creation and development of foundation models by streamlining, pre-processing and processing the data to assist in cleaning, transforming, and structuring raw data into usable formats. This helps to ensure that training data is representative and diverse, leading to better model performance. These processes are known as 'tooling' and play a crucial role in ensuring efficient, reproducible and responsible model development.

Foundation models continue to be an area of major innovation and development with new models regularly being released that outperform the previously best performing models (based on various benchmarks).[7] Foundation models are most established in text and image generation but are making very rapid progress in video generation and other areas.

In some cases, foundation models are also trained on domain-specific data: for example, BloombergGPT, a LLM for finance, was trained on financial domain-specific dataset constructed by Bloomberg and general purpose datasets.[8] This suggests that while for simplicity we have drawn a clear demarcation between the foundation layer and the application

6 CMA (2023). AI Foundation Models: Initial Report

7 See https://arxiv.org/abs/2009.03300 and https://blog.google/technology/ai/google-gemini-ai/#performance

8 https://arxiv.org/abs/2303.17564

layer described below, in practice this line can be blurred where a foundation model is conceived specifically for a sectoral domain (such as finance in the example above).

2.3 Application layer

The **application layer** refers to the last stage in the generative AI value chain where foundation models are integrated into apps and other user-facing products and services. It draws on both the compute and foundation layers and is crucial for generating practical use cases and implementations of generative AI. Building AI applications involves designing an orchestration framework that brings together an AI component (typically a foundation model) with front-end (i.e. user facing) and back-end (i.e. general operation) components. These applications map be standalone or integrated into existing products or services (such as AI assisted search).

Developing applications may also include 'fine tuning' of foundation models. Fine tuning is the process of taking a pretrained machine learning model and further training it on a smaller, targeted data set. The aim of fine-tuning is to maintain the original capabilities of a pretrained model while adapting it to suit more specialized use cases.[9]

Developers can also leverage and build upon existing foundation model applications to curate use case specific applications. For example, building an additional – more specified – layer on top of a general foundation model app such as ChatGPT. These are referred to as 'plug-ins' and allow for the development of multiple foundation model based applications across a variety of industries including education, hospitality, enterprise productivity software, marketing and others.

Another way that application developers commonly access foundation models for use in their apps is via APIs. These are services that allow a developer's app to interface with a foundation model. Many key foundation model developers provide access to their foundation models via APIs. In fact, the industry is witnessing the emergence of "model gardens", platforms offering a curated selection of models via APIs on a "model-as-a-service" basis. Examples include Amazon AWS Bedrock, Googles Cloud Services and Vertex AI, as well as Microsoft's Azure Machine Learning.

As well as access to foundation models, applications also require compute capacity to run. This is because a model's "inference" – each time the model is called upon by a user to make a prediction – also uses compute power. While the required compute power for a single inference is insignificant compared to the vast compute capacity required for training

9 While we define fine tuning as part of the application layer, fine tuning is sometimes considered to be part of the foundation layer. Whether it is considered to be part of the foundation layer or application layer does not significantly affect the framework applied in this report or our findings.

foundation models, when deployed at scale, the compute power required for an application can be substantial.[10]

Generative AI applications can be used in a wide range of industries and operational functions. Indeed, among companies recently surveyed by McKinsey across the globe, generative AI uptake ranged from 14% of businesses in the Energy and Materials sectors to 33% in Technology, Media, and Telecoms.[11]

Early examples of generative AI applications include, among many others:

- Conversational assistants including both general use (e.g. ChatGPT) or more specific applications;
- Coding assistants to support data scientists, data engineers and software developers, such as GitHub Copilot; and
- Image generation and editing tools within existing software (e.g. Adobe Photoshop) or as dedicated tools (e.g. Stable Diffusion or Firefly).

There is also emerging evidence on how generative AI applications are used in practice and on the impact their use can have. This includes, for example:

- A recent experiment showing that access to a generative-AI based conversational assistant increased the productivity of customer support agents, as measured by customer issues resolved per hour, by 14%.[12]
- An experiment showing that using ChatGPT raises the average productivity of mid-level professionals in writing tasks substantially. [13]
- The use of GPT-4 in management consulting tasks, shown to increase the speed of task completion and quality of outcomes among management consultants.[14]
- A trial of GitHub Copilot, a generative AI coding tool, showing that software developers who were given access to the Copilot completed tasks in a controlled experiment 55% faster than the control group.

2.4 Open versus closed source models

Foundation models used in the generative AI value chain can be either **open** or **closed source**. There are many differences between open and closed sources which pertain to how

10 We discuss this further in chapter 4.

11 McKinsey (2023). The state of AI in 2023: Generative AI's breakout year.

12 Brynjolfsson, E. et al (2023). Generative AI at work

13 Noy, S. et al (2023). Experimental evidence on the productivity effects of generative artificial intelligence

14 Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24-013).

they are pre-trained, fine-tuned, set for deployment, marketed and monetised. The sections below expand on the key differences in these two models.

Training

Open source refers to a collaborative approach, typically in software development and distribution, where the source code of a program or software is made freely available to the public. This means that anyone can use, modify, and distribute the code, often under a specific open-source license, as long as they adhere to certain conditions, such as providing attribution and sharing any modifications under the same open-source license. While there are numerous benefits to open-source models, their collaborative nature can leave them open to risks as individuals may seek to misuse the models, and they are generally harder to enforce governance policies on.[15] However, there have been various breakthroughs in the fine-tuning of open-source models which has made some comparable to fine-tuned closed-source models.[16]

Closed source models on the other hand, also known as proprietary software, typically refer to software that is developed and distributed with restrictions on access to its underlying source code. In the case of closed source software, the source code is not made available to the public; it is typically held and maintained by the organization or individuals who created the software. Users are usually granted a license to use the software, but they cannot view, modify, or distribute the source code. This allows the software developer or company to keep control over the code and protect their intellectual property. Closed source software may be subject to licensing fees or restrictions, and users are often limited in how they can use and customize the software. More generally, the best performing models currently in the market are closed-sourced. [17]

Routes to market and monetisation

Open-source and closed source models also have different routes to deployment, market and monetisation.[18] Developers of closed source foundation models, for instance GPT-4, release limited information on the training process, the data they have used, and the resulting weights. The data and weights used for pretraining and their models can only be accessed through APIs. The current routes to market for closed source models are:[19] (i) integrating foundation models into existing products/services, (ii) creating new products/services which can be monetised using subscriptions; and (iii) providing AI “as-a-service” which allows a third-party to use the model in its products/services.

15 AI Foundation Models Initial Report. (CMA 2023), page 41

16 AI Foundation Models Initial Report. (CMA 2023), page 41

17 AI Foundation Models Initial Report. (CMA 2023), page 48

18 AI Foundation Models Initial Report. (CMA 2023), page 15

19 AI Foundation Models Initial Report. (CMA 2023), page 15

Open-source foundation models on the other hand, provide information on both the data the model was trained on and the specific weights used. Retraining as well as finetuning open-source models is therefore possible. Open-source models can be marketed in two ways. Firstly, they can be marketed by AI development services which include pre-training/fine-tuning a model based on an open source model's framework or fine-tuning/ pre-existing pre-trained models and then providing the relevant third party with the ownership of the model including its weights.[20] Secondly, open-source models can be marketed by model hubs which is when third-parties develop and pre-train the model using the open-source model, and services from the developer can come in the form of support and infrastructure.[21] Lastly, developers of open-source models may also choose to release them with no means to monetise.

2.5 Overview of the interaction between different layers

While some companies such as Google, Nvidia, Amazon, Microsoft and Meta are active in many of the different layers of the generative AI stack, there are many companies active in only one layer or even just one subsegment of a layer (especially in the application layer). Additionally, there is a wide variety of different ways in which activities taking place in different layers may interact, especially given that products can be either "open" or "closed" source. Understanding these interactions and structures can be important in policy design as they can affect market incentives and the fulfilment of dependencies (see chapter 4) and challenges relating to coordination between different layers of the value chain (see section 5.1.3).

Figure 2 below provides a stylised representation of some of these different structures. For example:

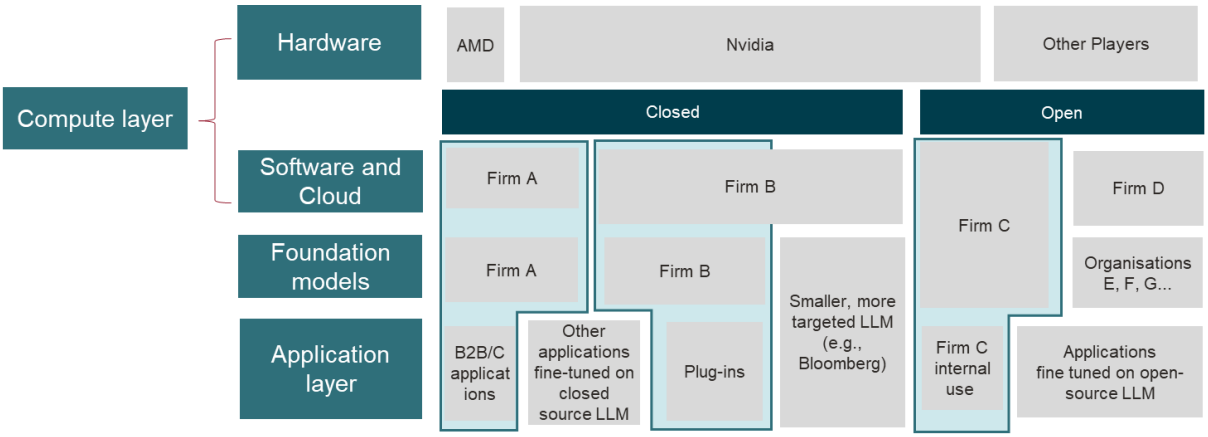
- Firm A is active throughout much of the generative AI value chain:
  - It has its own data centres and sells cloud computing services which it also uses for its own computing requirements;
  - It has developed a "closed source" foundation model with a chat-based interface. This is used by firm A in its own software products and can also be accessed by other organisations through an API. For example, a retailer could use the foundation model to provide generative AI-based sales assistance through its own mobile application.
- Firm C is also active throughout much of the generative AI value chain but takes a different approach to the compute layer and the use of its foundation model:
  - It has its own data centres, providing computing power that can be accessed through the internet (like firm A), but it only uses these data centres for its own operations;
  - It has developed an "open source" foundation model, which:

20 AI Foundation Models Initial Report. (CMA 2023), page 15

21 AI Foundation Models Initial Report. (CMA 2023), page 16

– Can be accessed by other organisations, which can fine-tune and deploy the model independently of firm C's own infrastructure, without having to start the model training from scratch;[22]
– Is used internally by firm C, for example to support its own software developers.

Figure 2 Stylised structure of the generative AI value chain



Source: Frontier Economics

Partnerships also play an important role in the structure of the generative AI value chain. Examples of partnerships include Open AI's partnership with Microsoft, and Amazon's partnership with Anthropic, amongst others.[23] Typically in these partnerships the more established firm focuses on the development of infrastructure-heavy compute capacity, where investments are often long-term and can be burdensome for a smaller player. Smaller firms can then focus investment in model and application development, while working closely with the larger firms to ensure new innovations and investments in the compute layer are aligned with progress in the foundation layer.

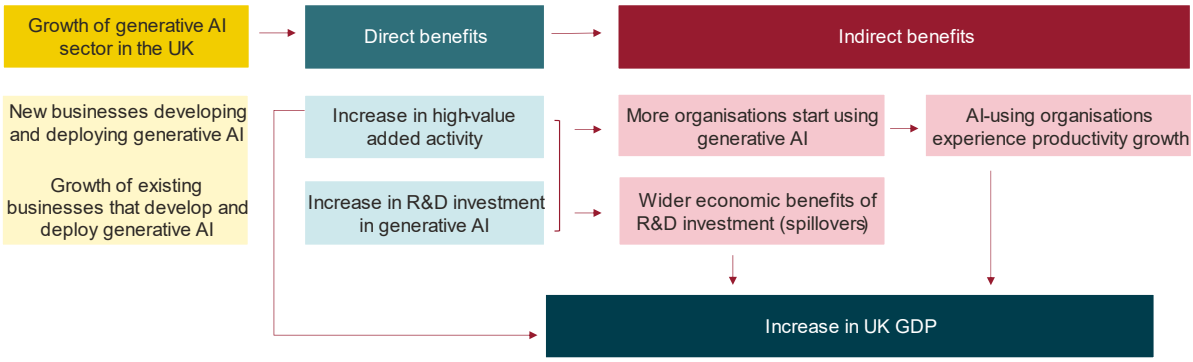
2.6 Economic benefits of the generative AI sector

Looking across all layers in the value chain, the creation of new businesses that develop and deploy generative AI and the growth of existing businesses operating in the sector could generate significant economic benefits. These benefits are illustrated in the figure below.

22 For instance, Meta's Llama 2 model is accessible to individuals, creators, researchers or businesses without a special license. Users receive the model code, the model weights, a user guide, a license and an acceptable use policy with the model download. It can be run locally on individuals' computers, although even for high end computers precision of the model would need to be reduced to do so.

23 See https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/ and https://www.aboutamazon.com/news/company-news/amazon-aws-anthropic-ai

Figure 3 Stylised structure of the direct and indirect benefits through increased generative AI uptake



Source: Frontier Economics

Direct benefits

The generative AI sector operates at the frontier of digital technology, and therefore it is likely to be a high-value added sector. Growth in a high-value add sector leads to overall growth in GDP if capital and labour move to this sector from lower value-added sectors. While it is too early to measure with any accuracy the value added per worker in generative AI, at a minimum we would expect productivity in the generative AI sector to be just as high as in other parts of the digital sector. This would be around £83k per worker per year which is higher than the UK average of £57k per worker per year.[24] However, because generative AI is a novel technology, the value added by each worker in the sector could be even higher. Our research suggests that the value added by workers in generative AI could be around 20% higher than in the ICT sector.[25]

Our calculations suggest that each additional 1,000 workers in generative AI development could contribute up to £16m of extra annual output to the UK economy[26]. Assuming that generative AI accounts currently for a small proportion of total AI employment[27], and using analyst predictions for the potential growth of the sector[28], it would be plausible that by 2033 more than 165,000 people in the UK could be working in the generative AI sector. This growth of the generative AI sector could generate an increase in UK annual GDP of around £2.6bn.

24 Statistics on average GVA per worker are derived from an OECD database and temporarily adjusted for 2023.

25 Lightcast (2022). Demand for AI Skills Triples in the UK Labour Market.

26 This is based on the assumption that these additional 1,000 workers would move from other comparable sectors to the generative AI producing sector. That is to say, we account for displacement of workers from other sectors.

27 Current AI employment is around 50,000 people in the UK. See: Perspective Economics (2023). Artificial Intelligence Sector Study, Research report for the Department for Science, Innovation & Technology (DSIT).

28 This is based on the assumption that the generative AI industry could grow at a CAGR of 42% over the next ten years. See: Bloomberg (2023). Generative AI to become a $1.3 trillion market by 2032.

Indirect benefits

Indirect benefits are the positive consequences from a larger generative AI sector that are not directly tied to but nevertheless are a result of it. These benefits can be significant and consist of:

- Impacts on other firms through innovation activity undertaken in the sector ("innovation spillovers"); and
- Impact of growth in the generative AI sector on adoption of generative AI.

Innovation spillovers

Investment in innovation taking place in one sector often leads to additional benefits in other parts of the economy. In the generative AI case, for example, new ideas and processes developed in the generative AI sector could be applied in other areas of AI research.

Using the ratio between R&D expenditures and annual operating expenses for the largest companies in the digital sector as a proxy, we assess what proportion of workers in the generative AI sector might work primarily on R&D (23%).[29] Assuming as above that by 2033 the UK generative AI sector could employ around 165,000 people, the value of innovation spillovers from the sector could be around £120m per year.[30]

Impact on adoption of generative AI

Adoption of generative AI is likely to account for a large proportion of the economic benefits of this technology. This is because, given the general-purpose nature of AI tools, take-up of generative AI could be very widespread, and the benefits for each adopter could be relatively large. As shown earlier in this chapter, emerging evidence suggests that using generative AI applications can have significant benefits on adopters in terms of time savings and greater quality of output produced.

Given the size of potential benefits from adoption, a key question in assessing the potential benefits of generative AI production for the UK is: would having more production of generative AI products taking place in the UK lead to greater adoption of these products among UK businesses?

To the best of our knowledge, the economic literature on the adoption of digital technology has not explored whether geographical areas where there is greater production of a new technology also experience greater adoption. Although investigating this link in depth was

[29] This is based on a calculation by Frontier which compares the annual R&D expenses with the annual operating expenses for the main companies working on AI and takes an average of them. For more information on companies R&D and operating expenses see: https://www.macrotrends.net/

[30] A recent study undertaken by Frontier Economics for the Department for Business and Trade has found that R&D investment typically produces social returns of 20% on top of the private returns from R&D investment. Assuming that this would hold for investment in R&D in the generative AI sector, we can assess the likely wider economic benefits (spillovers) that would result from R&D in the sector.

beyond the scope of this study, we conducted an econometric assessment using available data on European countries from Eurostat (including the UK). Because collection of data on AI adoption by national statistics authorities has only started recently, it was not possible to look at the adoption of AI specifically. Therefore, we investigate the relationship between production of Information and Communication Technology (ICT) services in a country and the take-up of cloud services in a country as a proxy of the possible relationship between generative AI production and AI adoption in a country. [31]

Our findings suggest that a doubling of the generative AI sector in the UK could lead to a doubling in the adoption of generative AI applications outside the sector, with associated further benefits of up to £20bn annually. The box below provides further detail on our calculations and on the strengths and limitations of this analysis.

[31] Due to limited data on AI production and adoption, we used cloud services as a proxy as it was the most closely related technology in which sufficient data was available.

Further detail on links between generative AI production and adoption

We use data on 33 European countries between 2014 and 2020 and run a fixed-effects regression that controls for the fact that some countries may have characteristics (such as their culture or institutions) that may have an effect on both the production of ICT services and the take-up of cloud services in the country. We find that a doubling of the share of people employed in the Service ICT sector is linked with a near-doubling of cloud usage in the non-ICT sector[32] of that country.[33]

While both the measure of technology production (the share of employees in a country that work in the ICT services sector) and the measure of technology adoption (the share of businesses in a country that use at least one type of cloud service) are not AI-specific, this analysis gives us an initial indication of whether a link between production and adoption of ICT technologies exists in the available data.

While it is possible that there are omitted variables driving both service ICT employment and the use of cloud services by enterprises over time, the econometrics is relatively robust evidence of a causal impact of ICT employment on adoption, and is stronger than a simple correlation. As AI production is a subset of the ICT sector, we use this evidence to assess the likely magnitude of the links between AI production and AI adoption. It is possible of course that in the case of AI the link between production and adoption might differ from other forms of ICT but we have no specific evidence to suggest whether that link may be weaker or stronger.

To estimate the potential impact of a doubling in generative AI adoption on productivity, we assume that, in line with our estimates for cloud services, a doubling in production leads to a near-doubling of adoption in the non-ICT sector. We then elaborate upon existing modelling estimates for the potential productivity uplifts of generative AI by sector, combing this with data on current usage of generative AI by sector, as well as data on gross value added across sectors to calculate potential resulting productivity gains and economic impact. Our calculations suggest that doubling the proportion of enterprises currently using AI in the UK could generate productivity gains of around £20bn across all industries per year.[34]

32 We look at the impact on cloud usage in the non-ICT sector specifically as this represents a more conservative approach due to cloud services usage by ICT sector businesses being more strongly correlated with ICT sector employment.

33 This estimate had a p-value of 0.055, meaning that there is a relatively low likelihood (5.5%) that the result is due to chance alone. It should also be noted that, due to data limitations, our econometric analysis is based on a linear regression while technology adoption regularly follows S-shaped curve.

34 These productivity gains are based on assumed productivity uplifts of up to 5% per industry. For information on the precise range of productivity uplifts per industry due to generative AI see: McKinsey (2023). The economic potential of generative AI: The next productivity frontier.

3 The UK's capabilities for generative AI production

Key findings

We discuss whether the UK has the capabilities required for businesses to compete effectively in the generative AI value chain. We consider direct and indirect measures of UK capabilities and seek to assess whether the UK is well placed to develop a comparative advantage in any of the layers of the generative AI stack. We find that:

- The UK currently has strong capabilities and competitive advantages relevant to the application layer (and, to a lesser extent, the foundation layer).
- UK capabilities relevant to the compute layer (especially the production of compute hardware) are currently weaker.
- There have been declines across several indicators of UK capabilities relevant to all layers, suggesting a general decrease in the UK's international competitiveness, which could lead to the UK struggling to sustain a competitive advantage in the generative AI stack in the future. In particular, we observe:
  - Evidence of an AI-skills gap in the workforce;
  - Declines in indicators related to the quality of the UK science base and AI research; and
  - A decline in the level of private investment in R&D and overall low international presence in patent-related indicators.
- It was not possible to fully map evidence into capabilities by layer. In addition, the current lack of information about certain capabilities limits further overall assessment of the UK's capabilities across the generative AI stack and in a given layer. As such, developing more precise measures of UK capabilities, potentially including new primary evidence, should be a priority area for future applications of this framework.

The first component of our framework is to assess the UK's current capabilities for participating in the generative AI value chain. In doing so, we seek to understand where the UK's capabilities are stronger, where they are weaker, and whether the UK currently has or could develop a comparative advantage in a particular layer of the generative AI value chain.

Since generative AI is an emerging and still rapidly developing market, there is some uncertainty about which capabilities will be most important for the development of generative AI and somewhat limited direct empirical evidence on current generative AI capabilities. We have therefore developed an approach to assessing UK capabilities that uses a range of both direct and indirect evidence. This is similar to the approach taken in other reports that have

looked at the AI capabilities of other countries (e.g. South Korea[35], Israel[36]). In particular, we assess:

- **AI skills in the workforce** – Participation in the AI value chain requires specialised skills. Assessing the UK's international ranking in workforce AI skills sheds initial light on how well the UK workforce is currently prepared for participation in the different aspects of the generative AI value chain.
- **The UK's academic innovation and research ecosystem** – The cutting-edge nature of generative AI means that participating in and benefiting from the generative AI value chain requires a high level of innovation that is rooted in research. We assess the UK's academic science base, particularly in AI, exploring the UK's capabilities in scientific discovery and innovation. Collaboration between academic and private sector researchers is also an important factor to assess, with potential benefits in the sharing of scientific advances for commercialisation, as well as increasing the UK's ability to attract talent and retain high skill students.
- **Indicators of private investment in innovation** – Private sector investment in innovation is important for commercialising viable innovations in the generative AI space. Current R&D investment levels and patent levels shed light on the willingness of UK businesses to invest in innovative technologies and their efficacy in commercialising these technologies.
- **Ability to attract VC investment into AI** – Given the innovative nature of generative AI, the development of new AI tools and products requires relatively high-risk investments without a guarantee of return. The private sector's ability (and, in particular, the ability of smaller firms) to attract VC investment is therefore an important factor in participation in the AI value chain.
- **Compute infrastructure capabilities** – Generative AI infrastructure mainly relates to the compute layer, which is an essential input into the foundation and application layers. Assessing the UK's capabilities to manufacture, assemble and access compute capacity is informative of its ability to participate in the generative AI value chain.
- **The UK's international position in related markets** – The computer services and R&D services sectors require similar skills (technological and software development skills) as would be required for the UK to participate in the development stage of generative AI across all layers. Assessing the UK's international competitiveness in those sectors therefore indicates if the UK could become an international leader in the generative AI value chain by leveraging those existing capabilities and advantages toward this new sector.

In the following subsections, we review the evidence on each of these areas in detail.

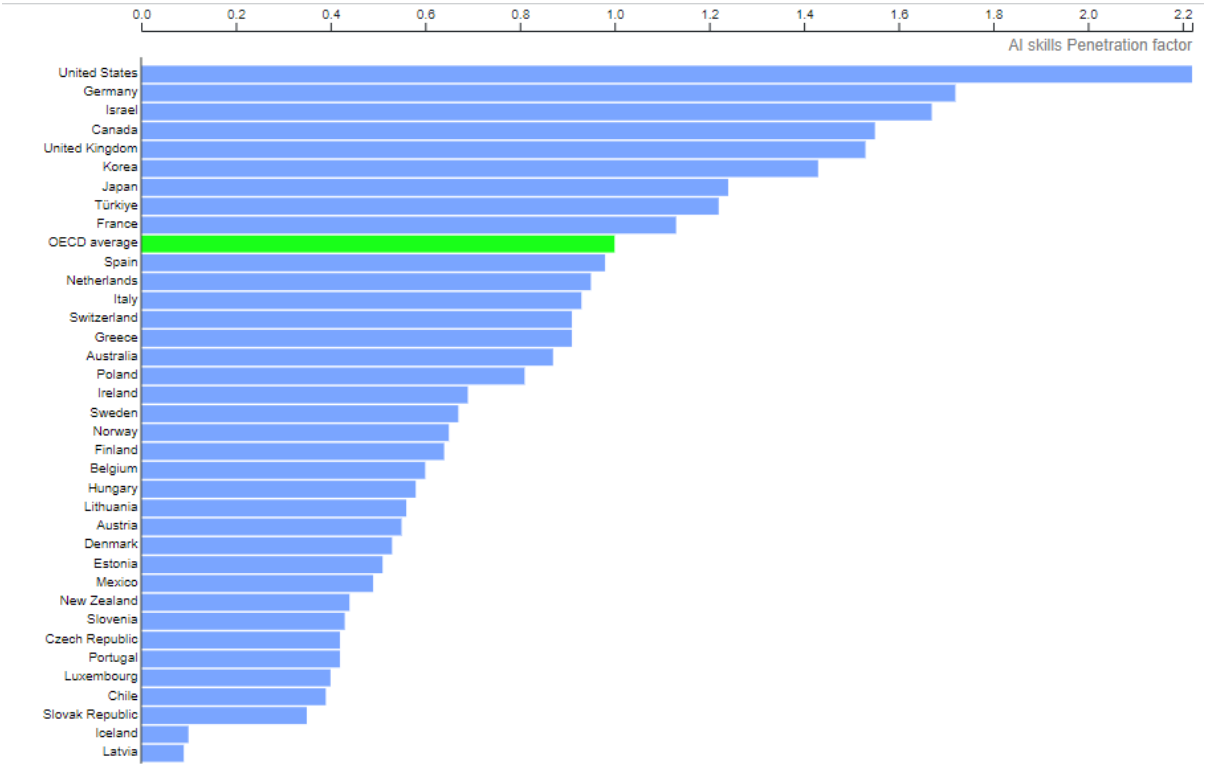
35 https://cset.georgetown.edu/publication/assessing-south-koreas-ai-ecosystem/

36 https://israel.um.dk/en/-/media/country-sites/israel-en/innovation-centre/state-of-ai-in-israel-2019-icdk-outlook.ashx

3.1 AI skills in the workforce

The UK already has a relatively good level of AI skills among its workforce. Figure 4 below shows the relative prevalence of AI skills amongst workers by country compared to the OECD average. This analysis is based on self-reported skills listed on LinkedIn between 2015-2022 and, as such, it is likely that this primarily identifies skills relevant to AI production, rather than skills relating to the use or adoption of AI tools. This is because individuals will typically list core skills that relate directly to their job and experience, and skills relating specifically to using AI, rather than developing AI, are a more recent phenomenon. That said, since the original study does not state what search terms were used for this exercise, it is not possible to definitively say which types of skills are captured. The indicator shows that the UK ranks fifth, with 1.5 relative penetration, meaning that UK workers are 1.5 times more likely than the average OECD worker to have AI-related skills. The UK is preceded only by the US, Germany, Israel and Canada on this indicator.[37] It is also in the top 10 when it comes to the prevalence of workers with AI skills in the technology, information and media sectors.[38]

Figure 4 AI skills penetration - OECD countries



Source: *OECD website https://oecd.ai/en/data?selectedArea=ai-jobs-and-skills&selectedVisualization=cross-country-ai-skills-penetration*

Note: *This chart shows the prevalence of workers with AI skills – as self-reported by LinkedIn members from 2015-2022 – by country and against a benchmark (either the OECD or G20 average). A country's AI skills penetration of 1.5 means that workers in that country are 1.5 more likely to report AI skills than workers in the benchmark.*

37 OECD.AI (2023), AI skills penetration.

38 OECD.AI (2023), AI skills penetration.

Average from 2015 to 2022 for a selection of countries with 100,000 LinkedIn members or more. The value represents the ratio between a country's and the benchmark's AI skills penetrations, controlling for occupations. Data downloads provide a snapshot in time. Caution is advised when comparing different versions of the data, as the AI-related concepts identified by the machine learning algorithm may evolve over time.

Even though the UK is among the leading countries for AI skills in the workforce, the majority of AI-developing businesses in the UK report skills gaps in a number of key areas.[39] In 2021, the Department for Digital, Culture, Media & Sport and the Office for Artificial intelligence released a study that reported a 50% skills gap in at least one of the following areas: (i) employees understanding of AI concepts and algorithms, (ii) programming skills and languages, (iii) software and systems engineering and, (iv) user experience.[40] This gap will be important to close in order to ensure that the UK can provide a sufficient supply of AI-skilled workers, given the increasing demand for those employees.

Moreover, a 2021 review of UK compute capacity found a shortage of large-scale computing professionals in the UK, ranging from system architects to system operation professionals and software engineers.[41]

Overall, the available evidence suggests that the UK currently has a good international position in terms of AI-skilled workers but current gaps in the required skills are significant (and demand for AI skills is likely to grow rapidly). However, there is a lack of more granular evidence about how those skills map to the various needs across the generative AI stack and it is not clear if the current UK AI-skilled workforce will be able to better support a particular part of the stack over another (one exception to this being clear evidence of a lack of large-scale computing professionals). A priority area for future applications of this framework should be to develop more precise measures of the skills relevant to each layer of the generative AI value chain and how the UK ranks internationally on these measures.

3.2 UK's academic innovation and research ecosystem

The UK has a strong overall academic science base. In its 2022 report about the international comparison of the UK research base, DCMS showed that the UK ranked third across numerous research indicators, including in the share of global publications, citations and highly-cited publications.[42] It has also ranked first in the world in its share of field-weighted citation impact.[43] However, the report also found that the UK's shares have decreased across all those indicators. As such, while the UK is still a strong leader in research, the continuation of those trends might impact the UK's competitive advantage in research in the future.

39 Office for AI, based on a survey of 73 firms who have employees working with AI or data science (2021).

40 https://www.gov.uk/government/publications/understanding-the-uk-ai-labour-market-2020/understanding-the-uk-ai-labour-market-2020-executive-summary

41 Government Office for Science (2021), Large Scale Computing: the case for greater UK coordination.

42 https://assets.publishing.service.gov.uk/media/628cd2828fa8f55615524e8c/international-comparison-uk-research-base-2022-accompanying-note.pdf

43 Field-weighted citation impact compares how a number of citations for a given set of publications compares to the average number of citations received by all world publications in the same field.

More specifically, the UK is also a leader in AI-related research. The top 10 leading European AI research universities are in the UK.[44] Data from the OECD shows that since 2021, the UK has been fourth in AI research publications, preceded only by China, the US and the EU block.[45] That said, the UK used to be ranked third in this statistic before 2021, after which it was overtaken by India, which kept the third place constantly in the last two years.[46] Therefore, as with the UK's position in research overall, while the UK is still a leader in AI research, there have been some slight declining trends in recent years.

Evidence also points to the UK having a relatively high level of cooperation between academia and the private sector. It is ranked fourth globally (if considering the EU as one region) in academic-corporate peer-reviewed AI publications (measured by academic journal citations).[47] This suggests that there is a focus in UK AI research both on achieving academic advances and also exploring applications and potential benefits of these advances in a commercial setting. This collaboration ecosystem between academia and the private sector is likely to be particularly important for generative AI innovation, especially in the application layer.

Having a strong UK academic base is also important in creating the supply of future AI skilled workers if it is able to attract talent and then retain this talent post-study. The UK attracts a large number of international students. In 2023, the UK was second after the US in attracting inbound postgraduates.[48] Past trends suggest that the UK can retain a significant proportion of the talent it educates. For example, 2022 research by the Department of Education showed that about 39% of EU and 15% of non-EU graduates had sustained employment in the UK five years after graduation. Those numbers might have changed in the past years due to immigration law changes, in particular with regards to EU graduates, but it shows the UK has the relevant academic international reputation to both attract and retain talent – both are important abilities in ensuring a relevant talent pool for future needs.

Overall, the evidence shows that the UK has the required academic and cross-academic-private sector collaborative relationships to foster the needed innovation, benefit from AI commercialisation, and provide talent to support future workforce needs. That said, across most of the evidence, declining trends suggest a possible loss of international competitiveness in this area. There is also some uncertainty about the relative importance of scientific research for different layers of the generative AI value chain and the relative importance of private sector collaboration in this research. Currently, there is a global trend of foundation model research occurring within private sector organisations, rather than universities, largely due to the compute requirements involved.[49] This may suggest that the UK's relative strengths in

44 EduRank (2023).

45 https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=ai-publications-time-series-by-country

46 https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=ai-publications-time-series-by-country

47 Stanford University, 2021. https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf

48 https://www.britishcouncil.org/sites/default/files/postgraduate_mobility_trends_2024-october-14.pdf

49 Benaich and Hogarth (2022), State of AI Report.

academic innovation and research are most relevant to the application layer, rather than foundation layer.

3.3 Indicators of private investment in innovation

Evidence shows that the UK has a good international position in private R&D, which can be considered as an indirect indicator of the local conditions for private innovation and commercialisation (e.g. the ability to source relevant skills, quality of regulatory regimes, and incentives for innovation). After recent revisions in the ONS methodology used to measure R&D spending, the UK's business investment in R&D (BERD) in digital sectors as a share of GDP has risen from middle-of-the-pack internationally to fourth place (around 0.53%), behind Israel, the US and Sweden.[50] This shows that the UK provides good market conditions for overall substantive market investment in innovation, which is important across all layers of the generative AI stack.

This private digital R&D investment is also focused on AI-related areas. A 2023 report for the House of Commons about R&D spending mentioned that in 2021, 23% of all private R&D investment was spent on computer programming and software development research. Both areas are relevant for the advancement of the generative AI stack as a whole and, in particular, to the application and foundation layers.

Looking at patent indicators, which is a relatively well-documented indicator of private investment outputs, shows the UK is somewhat behind other countries. When looking at the total number of patents, the UK ranked 13th globally, with only a 0.6% share in 2021. Even when accounting for the size of the UK compared to others, the UK was ranked 8th. A closer look at technology-related patents shows a similar view, with the UK having only 0.71% of the global share, ranking in 8th place.[51] Between 2010 and 2021, the UK was also 8th place internationally in AI patent applications and granted AI patents.[52] This suggests that, compared to international benchmarks, the UK is less likely to generate innovations that are taken to market through patenting. This evidence does not allow us to assess whether those trends would be more relevant to the development of software (i.e. mapped more to the foundation models and application layers) or to hardware (i.e. mapped better to the compute layer).

Overall, the evidence suggests that the UK has a rather good international position in indicators of private investment in R&D but a somewhat lower rate of patenting and commercialisation. Evidence that private digital R&D is focused on computer programming and software development suggests that current UK innovation may be more related to the application layer, but further evidence on this would be beneficial in future applications of this framework.

50 Frontier Economics analysis of ONS and OECD data.

51 https://www3.wipo.int/ipstats/key-search/search-result?type=KEY&key=221 https://www3.wipo.int/ipstats/key-search/search-result?type=KEY&key=221

52 Center for Security and Emerging Technology (2023), "Assessing South Korea's AI Ecosystem".

3.4 Ability to secure finance

Smaller start-ups with innovative ideas require access to venture capital (VC) and the ability to secure finance. Development of apps and, to an extent, foundation models are more likely than the compute layer, to be developed by smaller start-ups as the development costs are lower. As such, to ensure the UK is able to participate in the application and foundation layers, it is important to assess the level of VC funding that startups (particularly AI-related startups) can secure in the UK.

Firms focused on AI development have been able to raise substantial amounts of funding in the UK in recent years. In 2022, the UK was in the top 3 countries in AI VC funding, raising $7.0bn.[53] Forty percent of those funds were raised for companies in the financial and insurance services sectors and fifteen percent for companies in the healthcare, drugs and biotech sector.[54] Both of those sectors are also sectors where the UK already has an international advantage. In the four quarters before June 2023, "Financial Services" was the second largest sector of export from the UK after "Other Business Services".[55] More so, the UK contributed 18% to global exports of financial services, second only to the US.[56] Table 1 further shows that three of the 5 largest VC-funded AI startups in the UK were in financial and insurance services, and two were in healthcare. Out of those, only one relates more to infrastructure development, while the rest refer to software that would be more aligned with the application and foundation layers.

Table 1 Top funded AI startups in the UK as of December 2023

Company	Total raised (USD m)	Industry	Background
Checkout Ltd	1,832	Financial and insurance services	Develops a SaaS (software-as-a-service)-based platform that enables businesses to accept payments through in-country acquiring that uses Visa tokens for payment processing regardless of geographical location
Curve UK Ltd.	1,360	Financial and insurance services	Develops and operates an online financial technology platform that enables individuals to manage and track bank cards and finances in one mobile application

53 https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=top-ai-start-ups-per-country-and-industry

54 https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=top-ai-start-ups-per-country-and-industry

55 https://www.gov.uk/government/statistics/uk-trade-in-numbers/uk-trade-in-numbers-web-version

56 Frontier Economics based on WTO data

Company	Total raised (USD m)	Industry	Background
CMR Surgical Limited	1,234	Healthcare, drugs and biotechnology	Global surgical robotics business that develops Versius, a surgical robotic tool for healthcare providers
Babylon Healthcare Services Limited	993	Healthcare, drugs and biotechnology	Digital health company that combines AI and delivers access to healthcare including personalized health assessments, treatment advice, and face-to-face appointments through its mobile application
OakNorth UK Ltd	880	Financial and insurance services	Operates an AI-integrated platform that provides online banking solutions including personal saving accounts, loans, and business credit financing services

Source: OECD.AI https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=top-ai-start-ups-per-country-and-industry

The UK's current leadership in AI VC funding as a whole and in the sectors where it has particular international advantages, such as financial services, shows that market conditions and support for firms to participate in the generative AI sector are present. This is likely to be a particularly relevant strength for small AI businesses and start-ups developing AI applications in areas such as fintech, healthcare and professional services. However, it is less clear whether there will be adequate VC funding for AI applications in other sectors and this is something that will be important to monitor going forward.

3.5 Compute infrastructure

As described in section 2, for the purposes of this report we are interested in computing infrastructure in two ways: a) as an input into the development and deployment of foundation models and generative AI applications, and b) as a layer of the global generative AI value chain where the UK could in principle play an active role, through businesses operating in the UK for the purpose of providing computing goods and services to both UK and international customers. We consider these two aspects in turn here.

Availability of high-performance computing infrastructure in the UK

Compute infrastructure is an important building block in the generative AI stack as it is an input into the foundation model and application layers. It provides the high computing power that is required for the intensive data processing involved with the training of foundation models and the deployment of generative AI applications. Having good capabilities in accessing the compute layer will be important in having a strong and striving generative AI sector.

One way that access to the compute layer can be ensured is by having local compute infrastructure – i.e. UK-based supercomputers. Overall, the UK currently seems to have somewhat limited capabilities in terms of computing infrastructure. The UK has 14 of the world's top 500 most powerful computing systems. However, the most powerful of these

(ARCHER2) is only ranked 30th in the world. The UK's National AI strategy acknowledged that the UK's infrastructure, in regards to computing power, is lagging behind other countries such as the US, China, Germany and Japan.[57] It is also important to note that not all supercomputers can be used to develop and produce generative AI, but this evidence suggests that the UK does not have a good overall local ability in supercomputers.

To assess the UK's capabilities in computing power for AI in particular, we need to assess its ranking in supercomputers with parallel processing capabilities. One of the main inputs into an AI compute infrastructure are Graphic Processing Units (GPUs), which are chips that provide the parallel processing power needed for generative AI foundation model development and production. Only supercomputers with a large enough number of GPUs (or other equivalent parallel processing units) can be used to develop generative AI foundation models. The State of AI Report Compute Index from September 2023 shows that the UK has two supercomputers, ranking number 4 and 7, in the number of A100 GPU counts in the HPC cluster. That said, this analysis does not include other public and private computing solutions which might be using other GPUs (such as V100 GPUs).

Access to the relevant computing power can also be done through accessing global providers of data processing capabilities and data centres. For example, Amazon's AWS services, . Microsoft's Azure and Google's GPC service. Those services allow users to rent out processing power, which can be divided across data centres that those services have across the world. Publicly available data on the location of global data centre capacity is somewhat limited, however, one indicator is the number of servers installed by country. The UK currently has around 3.4 million servers installed, ranking fourth in the world behind the US, China and Germany. However, in per capita terms, the UK ranks 16th in the world behind countries such as Singapore, Ireland, the Netherlands, Latvia, Estonia, Finland, and Australia. Many of these countries have already established themselves as data centre hubs, with a comparative advantage in the provision of computer infrastructure as a service.

Overall, this suggests the UK has somewhat limited domestic compute capacity, ranking in the middle of the pack amongst advanced countries. We discuss evidence on the extent to which UK compute capacity represents a barrier to development of the generative AI sector in section 5.2.

The UK's performance in providing computing goods and services

Aside from being an important input into the foundation model and application layer, compute infrastructure is one of the layers of the generative AI value chain. As such, we also assess the UK's current capabilities for participation in this layer.

Evidence suggests that the UK has limited capabilities when it comes to manufacturing the inputs needed for the AI compute layer. GPUs and other processing components needed for the compute layer are heavily dependent on the semiconductor sector – a sector without high

[57] UK National AI Strategy, page 33.

capabilities in the UK. The government has acknowledged this in its recent report on the semiconductor industry in the UK, saying that "the UK cannot and should not aim to meet our semiconductor needs domestically".[58]

The supply of computing as a service is also a market with high barriers to entry due to significant economies of scale in the cloud market and a current shortage of GPUs.[59] The UK does have a developed data centres sector, and London is a major hub for data centres.[60] However, as discussed above, the UK ranks behind a number of other countries for server capacity per capita and it appears that there are constraints to the expansion of the sector, discussed in section 5.1.6 of this report.

Overall, this suggests that the UK is not well positioned to become a leader in the provision of compute hardware or compute as a service.

3.6 UK's comparative advantage in related sectors

Having an absolute advantage in a given sector means that the country has the relevant skills, access to required inputs and market conditions to create services with lower cost or higher quality than other countries. Having an existing international advantage in sectors which require similar skills and market conditions might indicate that the UK could also have a comparative advantage in the generative AI sector.

We have selected two sectors as being most relevant, at present, for assessing the potential comparative advantage of the UK in AI: the computer services and the R&D services sectors. Computer services require similar skills to those needed for generative AI development (technological ability, coding, problem-solving etc.) and R&D services are especially relevant due to the cutting-edge nature of generative AI and the need for innovation and R&D, particularly in the application and foundation layers.

Figure 5 below shows that in 2022, the UK was 6th in the world in terms of its contribution to the worldwide exports of computer services.[61]

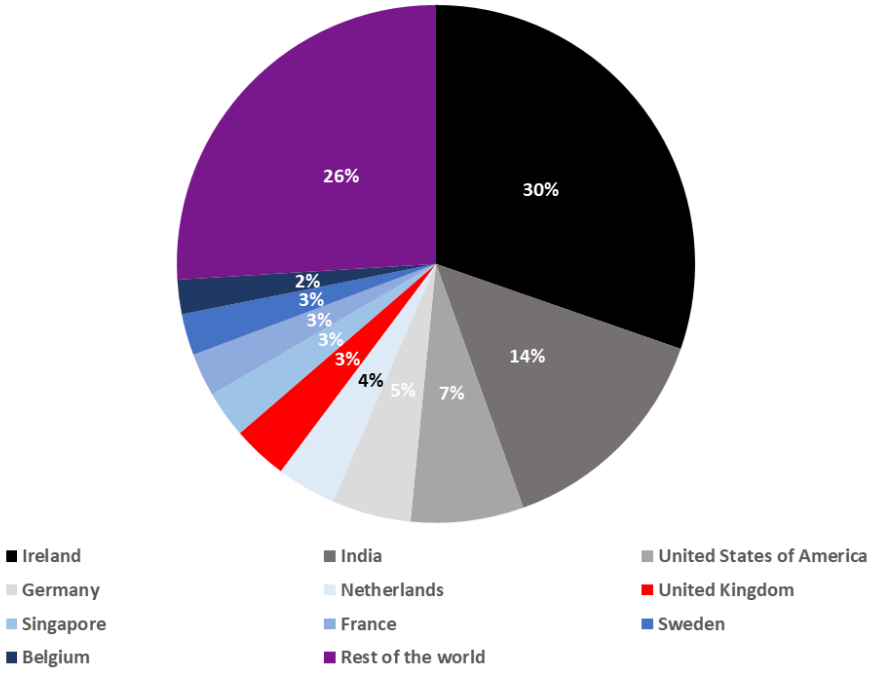
58 https://publications.parliament.uk/pa/cm5803/cmselect/cmbeis/1115/report.html

59 https://www.nytimes.com/2023/08/16/technology/ai-gpu-chips-shortage.html

60 TechUK (2020), "The UK Data Centre Sector"; CBRE, "Global Data Center Trends 2023".

61 https://stats.wto.org/

Figure 5 UK total exports of computing services out of global computer services exports – 2022



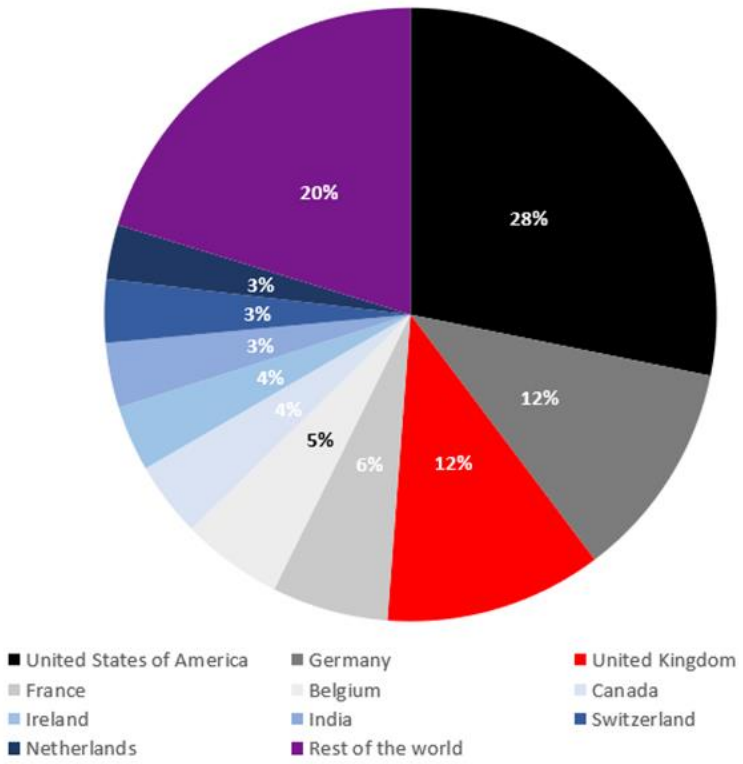
Source: Frontier Economics based on WTO data.
Note: This includes computer software and other computer services exports

Although exports of computer services only represent about 5% of UK exports, which is comparably low to other EU countries,[62] the chart above suggests that the UK has a significant global market in this sector – which can be leveraged toward the development of generative AI market that requires similar skills, capabilities and market conditions.

In terms of R&D services, Figure 6 below shows that in 2022, the UK ranked third in contributing to the global exports of R&D services.

62 In 2021, computer services (including other computer services) made up only 5% of UK services exports, which was significantly lower than the EU average (17%). Source: OECD.Stat, EBOPS 2010 - Trade in services by partner economy

Figure 6 UK total exports of R&D services out of global R&D services exports – 2022



Source: *Frontier Economics based on WTO data*
Note: *Includes Research and Development services exports.*

This suggests that the UK has a particular international advantage in R&D services (somewhat stronger than its advantage in computer services). Although R&D services only represents 4% of the UK's total service exports, its international competitiveness could potentially be leveraged towards the development of generative AI.[63]

Overall, the UK's position in computer services and R&D services provides indirect evidence to suggest that the UK may have the capabilities and market condition necessary to develop a comparative advantage in a generative AI value chain. The UK's somewhat stronger position in R&D services relative to computer services may suggest that its capabilities are more aligned with the application and foundation layers, rather than the compute layer.

3.7 Summary of findings

Given all of the evidence above, we assess that the UK has capabilities well aligned with generative AI development, particularly in the application layer. The UK has relatively good

63 Analysis of data provided by: OECD.Stat, EBOPS 2010 - Trade in services by partner economy

availability of AI skills, relatively high venture capital investment in applied technologies, and a strong science base and international reputation for research and innovation. Evidence also suggests that the application layer in particular can, and already does, leverage the UK's strong international standing in finance, healthcare and biotech to attract VC investment – key potential areas for generative AI applications. For the compute layer, evidence suggests that the UK does not currently have the relevant capabilities or international advantage in manufacturing inputs needed for the computing infrastructure, nor does it have any relative advantage in accessing the supply of those inputs.

Taken together, this evidence suggests that the UK is better placed to develop a comparative advantage in the development of AI applications and there may a case for additional government support for AI to be focused on the application layer in order to capitalise on these stronger capabilities. Table 2 below summarises the rationale and evidence behind these conclusions.

Table 2 Summary of the UK's capabilities in each layer

Layer in generative AI supply chain	UK capabilities	Reasoning
Application	✓✓✓	■ Availability of AI skills in the UK workforce ■ Relatively high VC investment in applied technology (fintech, biotech, health) in the UK ■ UK comparative advantage in financial services (a sector with the highest AI VC funding) can be a key area for developing and adopting AI applications. ■ Current comparative advantage in related sectors such as computer science and R&D services
Foundation	✓✓	■ Good international position of the UK science base ■ Availability of AI skills in the UK workforce ■ Current comparative advantage in related sectors such as computer science and R&D services
Compute	✓	■ Lack of current comparative advantage in the manufacture of computer hardware and limited presence in cloud computing sector

Evidence gaps and uncertainties

In this section, we have addressed the capabilities component of our framework, looking at UK capabilities for participation in the generative AI stack. However, there remain a number of key evidence gaps and unanswered questions that future applications of this framework and assessments of AI policy should seek to address/monitor:

- It was not always possible to closely map the evidence we found to a particular layer of the generative AI stack. Where possible, we attribute capabilities to a layer, but in the future, data about capabilities that map better to each layer can help assess the UK's participation in each layer of the stack.
- It is not clear which specific skills are currently or in future will be most important for each layer of the generative AI value chain. Future research should try to assess skills at a more disaggregated level, for example, separating out data science skills from data engineering and prompt engineering.
- Evidence about participation in the compute layer, in particular, was challenging. Information about all the relevant supercomputers and data centres that can provide generative AI computing power was not available, making it hard to assess the UK's capabilities and international standing in participating and accessing this part of the stack.

4 Interdependencies in the generative AI value chain

Key findings

We discuss the interdependencies at each layer of the generative AI stack and consider whether it is important for the UK to be active in certain layers of the generative AI value chain (for example, foundation models or compute) to ensure that other layers (for example, application developers) have access to the inputs they need to succeed. We find that:

- Compute capacity is a key input both for the pre-training of foundation models as well as for the deployment of generative AI applications.
- Compute capacity for the pre-training of foundation models could, in principle, be located anywhere in the world, however, data residency requirements may necessitate access to domestic compute capacity.
- Compute capacity to support deployment of generative AI applications may also require compute capacity located in the UK due to data residency requirements and to minimise latency, which could impact user experience of some AI applications.
- Access to cutting-edge foundation models is crucial for the development of applications. Currently there is good access to foundation models, including through open-source models, however, if foundation models continue to grow in size (such that only a relatively small number of large established players are actively producing foundation models) and open-source models (which currently provide developers with cost effective access to cutting-edge foundation models) do not keep pace with private models, there may be a case for supporting the foundation layer to ensure cost effective access.

In chapter 3 we have seen that the UK has significant capabilities that could be leveraged in the development of the UK's generative AI sector. However, there are some gaps in UK capabilities and the UK's capabilities are not equally distributed across all layers of the generative AI value chain. In particular, the UK has somewhat stronger capabilities relevant to the application layer of the generative AI value chain and may be well placed to develop a comparative advantage in the development of AI applications.

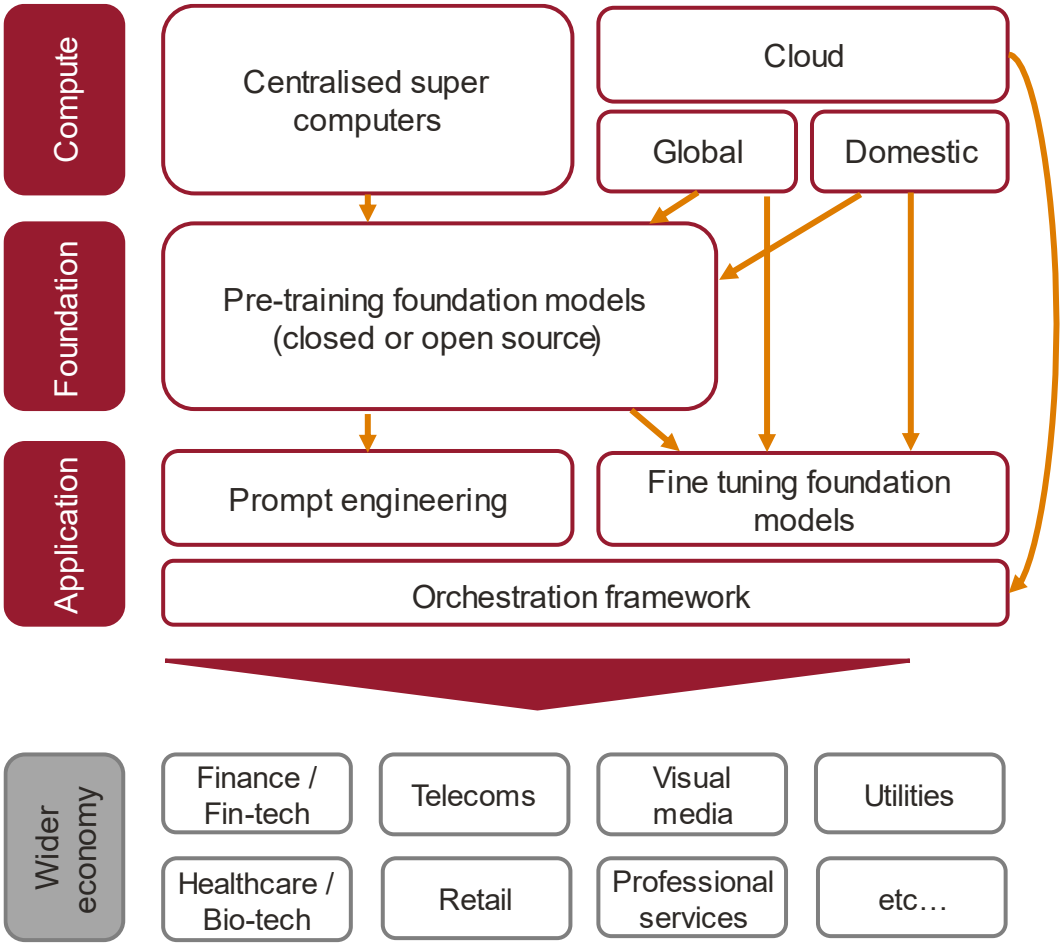
The next step in our policy framework involves assessing the interdependencies at each layer of the generative AI stack. We do this to understand whether it is important for the UK to be active in certain layers of the generative AI value chain to ensure that other layers have access to the inputs they need to succeed. In particular, we assess issues related to:

1. the current and likely future dependency of the foundation and application layers on compute capacity; and
2. the current and likely future dependency of the application layer on access to foundation models.

Our assessment draws on a rapid literature review and discussions with industry stakeholders.

Our understanding of the generative AI value chain at present is summarised in the figure below and described in detail in the rest of this section.

Figure 7 Illustration of the generative AI value chain



Source: Frontier Economics

4.1 Compute dependencies

As described in chapter 2, the compute layer provides crucial inputs to the other layers of the generative AI stack. These inputs include both computing hardware/infrastructure and specialist cloud computing software. This compute capacity is used in the foundation layer of the generative AI value chain for pre-training foundation models and in the application layer for fine-tuning models and deploying AI applications. As described in chapter 3, this compute capacity may come from either a specific supercomputer or from global cloud service providers, which allow users to rent out processing power, divided across data centres around the world.

Therefore, in general, "data centre capacity" is a subset of "compute capacity", as the latter includes both centralised supercomputers and distributed hardware. However, it is important to note that not all computers, whether centralised or distributed, are relevant for the generative AI stack. This because training and deploying generative AI models requires specific types of processing units.

Pre-training foundation models requires vast amounts of compute capacity, potentially thousands of specialised processing units (GPUs or TPUs) working for months at a time.[64] By contrast, fine-tuning of foundation models requires much less compute capacity – typically three or four GPUs working for half a week. However, in the longer term, it is possible that the proportion of compute capacity required for the deployment of generative AI applications exceeds pre-training of foundation models. While publicly available information on compute requirements is currently somewhat limited and difficult to compare between model training and deployment, one metric we can look at is estimated carbon emissions from the training and deployment of models. In the case of GPT-3 for instance (a model pretrained on old energy-inefficient hardware), the carbon emissions for pre-training the model were only about 5 times larger than the average monthly carbon emissions from the use of the model.[65] This means that the energy requirements of GPT-3 deployment has already exceeded the energy requirements of pre-training the model.[66]

Given these dependencies and the UK's moderate status amongst advanced countries for compute capacity (as identified in chapter 3 above), it is natural to ask whether there is a case for further government support and investment in domestic UK compute capacity in order to ensure sufficient supply.

In answering this question, it is important to consider whether *domestic* compute capacity is necessary or whether the foundation and application layers could rely on compute capacity located globally. Generally, there is no fundamental requirement for the compute infrastructure used in the generative AI stack to be located domestically or close to the developer/user. The training or deployment of models could, in principle, occur on compute infrastructure anywhere in the world. However, there are two factors that may necessitate domestic compute capacity:

- **Data residency requirements** – Because training an AI model involves potentially trillions of read and write operations, it is important that the training of the model occurs in the same location as where the model's training data is stored. As such, if the model's training data is subject to data residency requirements, this may place restrictions on

64 Towards Data Science (2023) Estimating the Cost of Training LLMs.

65 This is based on an assumption of about 10,000,000 queries per day. Please note that the comparison between pre-training and current model usage emissions does not factor in emissions from regular re-training. For more information see: Tomlinson, B. et al. (2023). The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans.

66 The compute requirements for generative AI applications can also depend on the size of the underlying foundation model. Therefore, suppliers often deploy smaller-sized models to answer less demanding prompts while larger models are queried for more complex requests. For example, Microsoft uses this approach to reduce latency in responding to Bing queries.

where the data can be stored and, in turn, restrict where the model can be trained. This is more likely to be the case for sensitive or personal data (for example, NHS data).

- **Latency issues** – Latency (i.e. delays in communication between developers of AI tools and compute infrastructure due to physical distance) is not generally a problem in the training of AI models. Latency delays are typically under a second, even when the compute infrastructure is in a very remote location relative to the developer, and these delays are unlikely to significantly impact the developer's experience. However, for the deployment of AI applications, latency may more substantially impact user experience. Domestic compute capacity with lower latency could improve responsiveness, speed, and user satisfaction.

The extent to which access to domestic compute is important for the foundation and application layers will therefore depend on the importance of data residency requirements and latency issues. Given that data residency requirements are relatively common in data sharing agreements and regulations, it is likely to be important for the UK to have a certain level of domestic compute capacity to support activity in other layers of the generative AI value chain.

Currently there is somewhat limited evidence on the extent to which access to domestic compute currently represents a barrier to growth of generative AI in the UK. We discuss this further in chapter 5 below.

4.2 Foundation model dependencies

There are two main potential sources of dependency between the foundation and application layers of the generative AI value chain. The first is the direct dependence of application developers on access to foundation models for developing generative AI applications. The second potential dependency that may arise is if being a credible global leader in application development (i.e. that is able to attract significant investment), also requires a certain level of activity in the foundation layer. This may be the case if knowledge spillovers and agglomeration benefits from the co-location of foundation and application development are very important for effective innovation and market leading development in both the application and foundation layers.

We are not aware of any existing evidence to suggest that knowledge spillovers and agglomeration benefits from geographic co-location are particularly important between the foundation and application layers of the generative AI value chain and therefore focus below on the more direct dependency of application developers on access to foundation models. However, in future applications of this framework, it may be appropriate to try to develop further evidence on the importance of agglomeration benefits between the different layers of the generative AI stack.

As described in chapter 2, foundation models are a crucial input to the development of generative AI applications. In particular, building AI applications involves designing an orchestration framework that brings together an AI component (typically a foundation model,

which may or may not be fine-tuned) with front-end (i.e. user facing) and back-end (i.e. general operation) components. Building applications does not necessarily require full access to the underlying foundation model – that is to say, it does not necessarily require access to the underlying weights (which is typically only possible for open-source models). For example, foundation models do not necessarily require fine-tuning (which would require access to the underlying weights), instead the application developer may use carefully designed prompts to elicit the desired responses, rather than fine-tuning the model.[67]

Currently AI developers and researchers in the UK have access to a range of foundation models, including open-source models. In its initial report on AI foundation models, the Competition and Markets Authority identified 160 foundation models, 68 of which were available through open access.[68] However, to our knowledge, no major foundation models have been trained in the UK, by UK-based organisations. It should also be noted that some open access models do not permit commercialisation of any applications based upon them.

In thinking about whether there is a case for supporting the foundation layer in the UK in order to ensure a secure supply of foundation models for application developers, it is important to recognise several key uncertainties in the future development of the market.

Continued growth in the size of foundation models

One key point of uncertainty relates to the future size and compute requirements of foundation models. To date, foundation models have grown progressively in size with larger models generally performing better, while the compute requirements for fine-tuning foundation models have fallen.[69] But there is some uncertainty around whether this will continue to be the case or whether foundation models will exhibit diminishing returns to scale at some point with additional pre-training data providing very little marginal value. If foundation models do continue to grow in size, this could have a number of implications for policy.

For example, if cutting-edge foundation models do continue to grow in size, this may create challenges for smaller businesses and start-ups to compete in this layer, and most activity and innovation may occur within a relatively small number of established larger businesses that have access to large amounts of compute capacity and can exploit economies of scale in production. In this context, if access to these models is crucial to the development of generative AI applications, then there may be an increased incentive for ensuring at least some of these foundation models are developed within the UK. For example, there may be a perceived risk that other countries developing cutting-edge foundation models could limit their accessibility in future.

---

67 Eric Horvitz (2023) The Power of Prompting.

68 Notable open access foundation models include Llama 2 (Meta), Stable Beluga 2 (Stability AI), and Falcon (TII UAE) - see AI Foundation Models Initial Report. (CMA 2023).

69 AI Foundation Models Initial Report. (CMA 2023).

At the same time, continued growth in the size of frontier models may mean that policies to support UK participation in the foundation layer aimed at start-ups or university spin-outs are relatively ineffective or disproportionally expensive (e.g. to purchase/subsidise adequate compute requirements). And policies may be more effective in generating UK participation in the foundation layer if targeted at making the UK an attractive place for larger established businesses to operate – for example, ensuring adequate availability of skills and providing a clear regulatory framework for AI safety and assurance.

However, it is also plausible that foundation models may not continue to grow in size. Future innovations may instead focus on producing more efficient, miniaturised models that can run locally on a user's phone or other device.[70] It may also be that AI developers do not require cutting-edge performance for future applications. For example, tasks like customer review classification or the generation of product descriptions can already be effectively executed using smaller models or ones fine-tuned for specific purposes. In the event that foundation models do not need to compete at the cutting edge to be useful in various applications, this could lead to smaller, fine-tuned closed-source models and open-source alternatives exerting effective competitive constraints on larger players. In this context, the risk of limited access to cutting edge foundation models would be lower.

Role of open-source models

A related area of uncertainty is the future role of open-source models. To date, open-source models have driven significant and rapid innovation by facilitating the creation of efficient, fine-tuned models that are publicly available and reduce the upfront costs for developers. This has provided alternatives in the market that are not reliant on expensive commercial models. However, the long-term competitiveness of open-source models is unclear.

As described in the Competition and Markets Authority's recent report on foundation models, open-source models face a number of challenges, such as:

1. Securing funding for their development;
2. Dependency on investor support;
3. The ability to commercialize some aspects of the models to be more attractive for investors;
4. Ethical concerns due to their potential misuse as their development relies on contributions from diverse sources;
5. Ensuring the continuous release of high-quality competitive models as some prominent contributors may discontinue their development efforts; and
6. Suppliers shifting away from open-source approaches once they have developed their ecosystem of partners and developers.

[70] For example, the latest Snapdragon mobile platform support generative AI models with up to 10 billion parameters on-device. See here.

The role of open-source models in the foundation model market will depend on how these challenges are addressed and how open-source suppliers adapt to changing market dynamics.[71] If open-source models continue to play a key role in the sector, this could provide UK businesses with widespread access to foundation models at a reasonable cost, allowing government to prioritise support in areas of UK relative strength, such as the application layer. As such, government may also want to consider whether there are particular policies that could support open-source models, addressing the challenges these models face.

Future advances and innovations

Another point of uncertainty relates to which layer of the generative AI value chain is likely to see the biggest technological advances and innovations going forward. In recent years, many of the biggest advances in the sector have been in the creation of foundation models. Going forward, it is difficult to predict whether we are likely to continue to see the biggest advances in generative AI coming from the foundation layer (such as, new approaches to pre-training or new data sources being utilised) or whether most innovation in the sector will now derive from advances in the application layer (such as, advances in fine-tuning and plug-in design).

While there is likely to be continued innovation across the whole of the value chain, where the biggest advances are likely to come from may impact future dependencies and influence the optimal design of policy. For example, if one believes that the key innovations in the foundation layer have already occurred and that future advances are likely to be in the application layer, then there may be increased justification for prioritising policies that support the UK's capabilities in the application layer. However, if one believes that there are still substantial advances likely to come from the foundation layer that could transform the sector and generate significant value, this may increase the justification for shoring up UK capabilities in the foundation layer, better equipping the UK to participate in and benefit from these advances, whilst also ensuring access to cutting edge foundation models for UK based application developers.

71 For further information see: CMA (2023). AI Foundation Models: Initial report, p50ff

Evidence gaps and uncertainties

In this section we have addressed the second factor in our policy framework on interdependencies between layers of the generative AI stack. We have identified a number of key unknowns in the future development of the sector that may impact future dependencies and the optimal design of policy. These will be important areas to monitor in future applications of this framework:

- The importance of UK based compute capacity, rather than relying on global cloud computing, either due to data residency requirements or, in the case of deploying AI applications, latency issues.
- The future size and compute requirements of foundation models and whether foundation models exhibit diminishing returns to scale at a certain point or can be trained more efficiently.
- The future role of open-source foundation models in both research and commercial development and whether open-source models continue to provide efficient access to foundation models for fine-tuning to specific applications.
- The importance of geographic co-location and agglomeration benefits between the different layers of the generative AI stack.

5 Potential barriers to the growth of generative AI in the UK

Key findings

We discuss the reasons why the private sector may not be able to fully reap the potential benefits from generative AI (or mitigate potential risks) without government intervention. We find that:

- Government support for the sector may be justified by:
  - the potential for innovation in the sector to benefit a wide range of people (beyond the companies investing in innovation in the sector);
  - capital market imperfections that reduce the availability of finance for the sector;
  - coordination issues between different layers of the value chain;
  - potential underinvestment in AI skills; and
  - safety and security concerns.
- Government should also consider whether there are barriers created by current policy in terms of: planning regulation and constraints on data centres; use of generative AI products in public sector organisations; and lack of clarity around AI regulation.
- While these growth barriers are relevant to all aspects of generative AI production, their relative importance is likely to differ between different layers of the generative AI value chain, as summarised in the table below.

Layer in generative AI supply chain	Potential role for Government	Importance of skills and quality of science base	Importance of barriers to access to finance for new ventures	Importance of barriers to investment in innovation	Importance of risks from AI use potentially limiting production/ adoption	Importance of digital infrastructure (connectivity, data centres)
Computing	✓✓	✓✓	✓	✓✓	✓	✓✓✓
Foundation	✓✓	✓✓✓	✓✓	✓✓✓	✓✓	✓✓✓
Application	✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓

5.1 Reasons why there might be a need for government intervention

So far in this report we have assessed the first two factors of our framework (Capabilities and Dependencies). In assessing the UK's capabilities, we identified that relatively strong capabilities in the application layer mean that the UK is well placed to develop a comparative advantage in generative AI applications (i.e. become a leading producer and exporter of AI applications and services). As such, there may be a case for additional government support to prioritise the application layer, capitalising on this opportunity.

In assessing dependencies in the generative AI value chain, we identified that, depending on factors such as data residency requirements and the role of open source models, it may be important for the UK to also be active in the foundation and compute layers of the generative AI value chain to ensure that application developers have access to the inputs they need to succeed.

The next step in our policy framework involves assessing whether there are barriers or opportunities that Government could address to foster the growth of the generative AI sector. Specifically, applying our framework involves asking whether there is a role for policymakers to play in:

1. ensuring that existing UK capabilities are maintained, at a minimum, or improved upon, where this is feasible/cost-effective; and
2. ensuring that the private sector is able to successfully leverage these capabilities.

To understand this, before jumping to potential policy options, it is important to assess the rationale for government intervention in the generative AI sector and whether there are barriers or opportunities that government could address. Indeed, there are a number of reasons why the private sector may not be able to fully realise the full potential economic and social benefits from generative AI (or mitigate potential risks) without government intervention. This includes economic reasons such as externalities, coordination problems, and capital market imperfections, but may also include constraints from existing policies affecting the sector. We describe these reasons below. In section 5.2 we assess the relevance of these for each of the different layers of the generative AI stack. In chapter 6, we consider what policy options could potentially address these barriers.

5.1.1 Positive externalities

As described in section 2.6, the production of generative AI is a highly innovative activity and, in addition to direct economic benefits, generative AI production occurring in the UK may generate substantial indirect benefits both through R&D knowledge spillovers and through impacting the rate of AI adoption by businesses in the UK. Because these indirect benefits are not fully reflected in the market incentives for private sector businesses investing in AI, the amount of private sector investment is likely to be less than optimal from a social point of view. As such, there is a case for government to support and invest in generative AI production to raise the total level of investment. Indeed, R&D spillovers are part of the rationale for the UK government's current innovation and R&D support initiatives more generally. Studies in the US suggest that social returns to R&D due to knowledge spill-overs are over three times as large as private returns.[72] A separate question is to what extent government investment in innovation should prioritise AI over and above other technology areas. While this question is beyond the scope of this report, the scale of the potential indirect benefits from AI production

[72] See Bloom, Schankerman, and Van Reenen (2013), Identifying Technology Spillovers and Product Market Rivalry, Econometrica 81 (4): 1347–93; and Lucking, Bloom, and Van Reenen (2020), Have R&D Spillovers Declined in the 21st Century?, Fiscal Studies 40 (4): 561–90.

(especially the potential impact on faster adoption of AI across the economy) suggests that there may be a strong case for making AI a focus of government innovation support.

5.1.2 Capital market imperfections

All capital markets are subject to information asymmetries between borrowers and lenders. That is to say, lenders cannot perfectly observe the borrower's reliability and likelihood of successful repayment. Nor can lenders perfectly monitor the borrower's actions and whether they are maximising their likelihood of successful repayment. These information asymmetries increase the cost of borrowing for businesses and act to reduce the total level of investment in the economy.

While capital market imperfections exist in all sectors, it is possible that the emerging generative AI sector is particularly susceptible to these issues. Indeed, there is considerable evidence that information asymmetries and the cost of borrowing are higher for small innovative businesses.[73] With the rapid pace of development and technological complexity in generative AI, it may be difficult for investors to accurately assess the potential and capability of generative AI businesses – which are often small start-ups. Investors may also find it more costly to understand and monitor the actions of the generative AI businesses in which they invest. These heightened informational challenges can act to increase risk and costs for investors, leading to higher capital costs for businesses. As a result, there could be insufficient funding from financial institutions for innovative activities such as AI research and development of new AI products and services.

5.1.3 Coordination problems

In many cases, a successful outcome depends on the coordinated action of several parties. In these cases, private sector institutions may require sufficient confidence or assurance that other parties will take certain actions or investments before choosing to invest themselves. With generative AI, such coordination problems could exist between different parts of the generative AI value chain. For example, developers of AI applications may be worried that producers of foundation models may withdraw access, or producers of foundation models who intend to charge developers for access may under-invest in the development of new models because they are not sure how much demand there will be for their model. There may also be coordination problems in terms of the interoperability of different models and applications. Coordination problems could also exist between compute providers and developers in the foundation and application layers, due to uncertainty about how much data centre capacity will be required and the potentially considerable lead times involved in scaling data centre capacity.

[73] Hall and Lerner (2010), The Financing of R&D and Innovation, Handbook of the Economics of Innovation.

5.1.4 Underinvestment in skills

There is evidence to suggest that there has been a substantial and long-term decline in the overall volume of employer training and investment in training in the UK.[74] Underinvestment in skills can occur for a variety of reasons. Employers may not provide sufficient training because it can be difficult to predict the benefits of that investment or because they may not fully realise the benefits from that investment (for instance, the employee may move away). Similarly, employees may underinvest in their own training when they cannot credibly convey or demonstrate their true skills and abilities to their employers.

These issues are present across all sectors, but they could be particularly relevant to the generative AI sector in two ways. First, underinvestment in the skills required by businesses to adopt and use generative AI applications could limit the demand for generative AI applications and therefore the growth of the sector. Underinvestment in such skills may be likely due to considerable uncertainty about the future development of the sector and exactly what skills will be beneficial or required for businesses using generative AI applications.

Second, cutting edge skills are crucial for the development of generative AI and there could be underinvestment in these skills due to the fast-evolving nature of the sector causing uncertainty about which skills will be most valuable and creating difficulty in demonstrating these skills to employers (e.g. without dedicated qualifications having been developed yet). This said, the high value already placed by businesses on AI skills provides substantial incentive for businesses and individuals to invest in these skills, and this is likely to (at least partially) offset the potential barriers described.

5.1.5 Safety and security

The development and use of generative AI presents a number of potentially significant risks to people's safety, mental health, privacy, human rights, and intellectual property.[75] While organisations that develop and deploy generative AI have significant commercial incentives to invest in the safety of their products, as in all sectors where there is significant potential for risk – such as food production and road safety – private investments can be effectively complemented by government and third sector involvement.

It is beyond the scope of this report to explore fully the potential for AI safety risks and the ways in which policymakers can help prevent and mitigate them. However it is important to recognise the role that safety and security could play in influencing the development of the UK's generative AI sector:

- A **lower likelihood of risks** to safety from the use of generative AI products and services will **increase user demand** for these products and services. Greater demand would in

[74] Chartered Institute for Personnel and Development, Addressing Employer Underinvestment in Training, July 2019.

[75] AI regulation: a pro-innovation approach, DSIT, March 2023.

turn make the development and deployment of generative AI in the UK more attractive to entrepreneurs, businesses and investors.

- **Certainty over responsibility and liability** for potential harms could also enhance user demand and support business decisions to invest in the development and/or deployment of generative AI in the UK.

- Over and above the factors above, justified **trust** in AI systems is necessary for potential users (consumers and businesses) to want to use generative AI products and services, and also to feel comfortable sharing data for the purpose of pre-training and fine-tuning models (which is particularly important given the crucial role of data as an input to these models).

5.1.6 Other policy issues

There are also a number of policy issues that may be limiting or could in future limit the UK's ability to grow and benefit from generative AI:

Planning restrictions and constraints on data centres

Generative AI requires significant data centre resources for both the foundation and application layers. As such, growth in the generative AI sector will likely require significant growth in UK data centre capacity.[76] It is therefore important to assess whether planning law and the availability of renewable energy for data centres is sufficient to support the necessary growth in data centre capacity or whether there are significant barriers to expansion.

In a number of cases, European data centre hubs are facing direct restrictions from local governments on expansion of data centre capacity due to concerns around the availability of land and power. For instance, local governments in Dublin, Frankfurt and Paris have stopped new development of data centres in and around these cities[77], while the Dutch government has recently restricted the development of new hyperscale data centres as it deemed the scarcity of space and energy as too great.[78] While the UK does not impose direct restrictions on new data centres, the cost and complexity of planning regulations and the availability of renewable power (a key concern for many hyperscale data centre providers) could act to limit investment and growth of the sector.

[76] Tier 3 data centres, which are the most suitable for AI applications due to features like onsite assistance, power or cooling redundancy, were operating at around 1,500 MW in 2022 and their capacity is expected to increase to around 3,300 in 2029 at a CAGR of 11% See: Mordor Intelligence (2023). UK Data Centre Market Size & Share Analysis - Growth Trends & Forecasts (2023 - 2028).

[77] Cushman & Wakefield (2023). Data center market comparison.

[78] Savills (2022). European Data Centres, Deep dive in the data sphere.

Constraints to demand for generative AI products in public sector organisations

In many new and innovative sectors, public sector organisations can act as early adopters, providing a secure source of demand and helping the nascent sector commercialise and grow. In the case of AI, there is a clear potential for public sector organisations to play such a role, with many public organisations processing large volumes of routine tasks and potentially benefiting substantially from the application of generative AI technologies. Indeed, the NHS already uses AI tools to, among other things, analyse X-ray images and help clinicians read brain scans more quickly.[79] However, widespread generative AI adoption in public organisations depends on careful management, oversight and regulation of AI safety risks. Until a clear regulatory framework exists that generates trust in the safe application of AI technologies to often sensitive areas overseen by public sector organisations, public sector demand for generative AI is likely to be constrained.

Existing regulations (or lack of clarity around regulations) constraining businesses ability to produce or adopt AI products

Regulatory incoherence has the potential to stifle competition and innovation in generative AI by causing businesses and start-ups to leave the market. Poorly designed regulation may result in businesses having to spend excessive time and money complying with complex rules instead of creating new technologies, and this is likely to disproportionately affect smaller businesses and start-ups.[80] As such, the UK government has already stated an intention to create a clear and unified approach to regulation with a cross-cutting, principles-based approach.[81] As the sector and this regulatory framework continue to develop, it will be important to monitor and assess the efficacy of this regulatory approach and whether it is adequately fulfilling its dual aims of promoting trust and facilitating innovation.

Access to data

Data is a crucial input for generative AI, both in the pre-training of foundation models (which requires vast quantities of generic text, audio or image data) and in the fine tuning of models to specific applications (which requires specific datasets relevant to the intended application). The quantity and quality of available data directly affects the performance, accuracy, and reliability of generative AI models. Available datasets are often not prepared with training of AI models in mind. For supervised learning, datasets also need to be labelled, a task which is typically costly, repetitive and time-consuming. Access to data was identified as the most common barrier to commercialisation experienced by AI related projects receiving UKRI

79 NHA (2023). Artificial Intelligence, guidance for patients and service users.

80 Evidence to support the analysis of impacts for AI governance, Frontier Economics, 2023.

81 AI regulation: a pro-innovation approach, DSIT, March 2023.

funding.[82] A wide range of policies can affect data access, including intellectual property / copyright law, and government data openness policies. We discuss this further in chapter 6.

5.2 Importance of these barriers for each layer of the generative AI value chain

While all of the potential barriers identified above have relevance to all layers of the generative AI value chain, certain barriers are likely to be more important for particular layers of the generative AI stack. There are also differences in how these barriers manifest in each layer of the value chain. We discuss some of these key differences between layers below.

Importance of skills and quality of science base

Availability of skills and the quality of the science base are of key importance to all layers of the generative AI value chain. Particularly for innovation activities, cutting edge skills and R&D capabilities are essential. However, there are likely to be differences between the layers in terms of the types of skills and knowledge required. While the compute and foundation layers are most likely to continue to require workers with a strong background in computer science and mathematics, the skills required in the application layer may be more varied.

For example, a key role in the generative AI application layer is prompt engineering.[83] This involves designing inputs for generative AI tools that will produce optimal outputs. While a prompt engineer may benefit from a background understanding in computer science, the core skills of the role also depend on linguistics, philosophy, and psychology. Additionally, as activity in the application layer is closer to the consumer facing end of the value chain, the application layer will also depend more on broader skills such as product design, market research and marketing. Successful application of generative AI tools to specific sectors, such as finance and healthcare, will also require sector specific knowledge, skills and experience.

As generative AI continues to evolve rapidly, the skills needed to support the sector are also likely to evolve, particularly in the application layer. As such, it will be crucial to continue monitoring whether there are key skills barriers, and not just within the traditional focus areas of computer science, data science and mathematics.

Importance of barriers to finance for new ventures and innovative activity

As mentioned above, all layers of the generative AI stack may face constraints from access to finance. However, this is likely to be more pronounced in the foundation and application layers. This is for a number of reasons. Firstly, due to the lower initial compute and infrastructure requirements in the application layer, there is likely to be a greater number of small players and start-ups than in other layers, and these smaller businesses will likely face greater

82 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021.

83 What is prompt engineering? McKinsey, September 2023.

challenges securing finance due to more pronounced information asymmetries with lenders and investors.

Secondly, the rapid pace of development in the sector may make it difficult for investors to assess the true potential and capability of generative AI businesses in the foundation and application layers, further exacerbating information asymmetries.

Thirdly, the nature of innovation activity in the application layer is somewhat different due to being more customer facing. In particular, it involves a broader set of innovation activities and intangible investments in product design and market research, which can be difficult to value and finance. It will therefore be important to monitor going forward whether there is sufficient financing of intangible investment in AI more broadly, not just for the training of new models.

Importance of safety risks from AI use potentially limiting production/adoption

AI safety risks are important for all layers of the generative AI value chain, however, we have identified several ways in which the appropriate management of AI safety is likely to be a particularly relevant issue closer to the application end of the generative AI value chain. Firstly, from a data perspective, the data used in fine tuning of generative AI models is more likely to include personal data and data held by individuals or commercial organisations, compared with pre-training foundation models (which generally depends on vast corpuses of publicly available data).[84] As such, the application layer depends more directly on clear regulation of the use of personal data for commercial purposes and the willingness of organisations and individuals to share sensitive data (which depends on having trust in the safe application of AI technology).

Secondly, by the nature of the consumer facing application of AI technologies, the application layer is more directly exposed to risks related to bias and safety. Indeed, the UK governments proposed framework for AI regulation focuses on the outcomes of the application of AI, rather than direct regulation of AI technologies.[85] As such, the application layer will benefit most from coherent and clear regulation.

Thirdly, as discussed in chapter 3, the UK may be especially well-placed for the development of generative AI applications in financial, professional and legal services, as well as health and biotechnology. All of these areas (especially financial services and healthcare) are areas where assuring the safety and security of AI applications will be particularly important, due to the highly sensitive nature of relevant data and outcomes.

Importance of digital infrastructure (connectivity, data centres)

While adequate broadband connectivity and data centre capacity will be necessary for the deployment of generative AI applications, digital infrastructure is likely to be particularly

84 This said, we note that there is considerable uncertainty about whether this will continue to be the case for pre-training of foundation models going forward and what data can be used for pre-training is an open issue.

85 AI regulation: a pro-innovation approach, DSIT, March 2023.

important for the compute and foundation layers (at least in the short term). Pre-training of foundation models requires substantial compute resources, generally distributed across multiple high performance data centres. This said, in the longer term, as discussed in chapter 4, it is possible that the proportion of compute capacity required for application and deployment of generative AI technologies exceeds pre-training of foundation models. We also note that the nature of data centres required for application of AI technologies will be different to that required for pre-training. For example, real time, low-latency application of generative AI may require more decentralised 'edge' data centres, rather than centralised supercomputers in hyperscale data centres. It will therefore be important to continue to monitor the current and likely future relative computation demands of different layers of the generative AI value chain and whether there are barriers to the availability of appropriate data centres and infrastructure. Due to lead times in establishing new data centres and scaling up domestic compute capacity, it will be important to identify any issues as early as possible.

Empirical evidence on the extent to which compute capacity is currently a barrier to growth of the generative AI sector is limited. One survey of mostly US based AI researchers by the Center for Security and Emerging Technology (CSET) found evidence that AI researchers are not primarily or exclusively constrained by compute access, with the availability of AI skills being a more commonly reported barrier.[86] They also found that AI researchers in academia are no more likely than researchers in industry to report lack of compute as a barrier to their research, that most researchers select grant funding as the resource that would be useful to them (though a considerable number also select compute), and that some researchers express concerns about an exclusive focus by government on scaling up compute.

A separate global survey of researchers and scientists by Nature found similar results: approximately 50% of researchers and scientists feel there are barriers preventing them from developing or using AI, with the most commonly cited barriers being lack of skills and a lack of funding (though a considerable number also mention lack of compute).[87] If we narrow this survey to only the 56 complete responses from UK based researchers, we find that 60% feel there are barriers preventing them from developing or using AI, and the most common barrier cited is a lack of skills. However, 53% of those who report barriers to them developing or using AI cite availability of compute as one of the barriers they face, and this goes up to 70% if we consider only responses from researchers who's work directly involves studying or developing AI.

As discussed in chapter 3, the UK currently ranks in the middle of the pack among advanced countries in terms of high-performance computing capacity. However, the UK government has committed to substantial investment in this area (see chapter 6) and Microsoft have recently announced £2.5bn of investment in UK compute over the next 3 years.[88] To ensure the

86 Micah Musser, Rebecca Gelles, Ronnie Kinoshita, Catherine Aiken and Andrew Lohn, "'The Main Resource is the Human'" (Center for Security and Emerging Technology, April 2023). https://doi.org/10.51593/20210071.

87 Richard Van Noorden and Jeffrey M. Perkel (2023), "AI and science: what 1,600 researchers think", https://www.nature.com/articles/d41586-023-02980-0

88 https://www.gov.uk/government/news/boost-for-uk-ai-as-microsoft-unveils-25-billion-investment

success of public and private investment in compute capacity in the UK, it will be important to monitor planning law and renewable energy availability, to ensure these do not unduly limit data centre expansion.

Evidence gaps and uncertainties

In this chapter we have addressed the third factor in of our policy framework, looking at reasons why the private sector may not be able to fully reap the potential benefits from generative AI (or mitigate potential risks) without government support. However, there remain a number of key evidence gaps and uncertainties that future applications of this framework and assessments of AI policy should seek to address/monitor:

- As generative AI continues to evolve rapidly, the skills needed to support the sector are also likely to evolve, particularly in the application layer. It is therefore important to monitor what the key skill areas are and whether there are skills gaps or barriers to skill acquisition in these areas, noting that these skills may be outside the traditional focus areas of computer science, data science and mathematics.
- While we have identified a number of potential barriers to growth that may justify government support, additional evidence on the magnitude and relative importance of these barriers would be helpful in guiding policy. Such evidence could come from surveying generative AI businesses or quantitative analysis of AI investments and VC funding in the UK.
- There is also considerable uncertainty over future compute requirements of the sector and a fuller assessment of UK data centre capacity (both current and forecast) could help identify the importance of barriers to growth in this area.

6 Policy options to support generative AI

Key findings

In this chapter, we assess the policy options available to the UK to support the development of the generative AI sector. Our initial analysis finds that existing evidence of effectiveness is strongest for:

- Policies that aim to support private sector innovation through loans or grants, especially when provided to SMEs;
- Policies to attract high-skilled talent from abroad;
- Policies to develop a homegrown talent base, through funding graduate and postgraduate education in STEM and AI-related subjects; and
- Supporting start-ups and scale-ups with access to finance.

Support to SMEs and to start-ups and scale-ups is likely to be more relevant for the foundation and application layers, compared to the compute layer, due to the large economies of scale in compute which may make it challenging for smaller enterprises to compete.

Initiatives to promote the safety of AI use, improve relevant infrastructure, and appropriately manage access to data could also be effective but there is not as much evidence on their impact from existing studies and, as such, it may be appropriate to prioritise impact evaluations in these policy areas.

6.1 Our framework to assess policy options

We have seen in chapter 5 that there are several potential barriers to the growth of the generative AI sector and opportunities for UK policymakers to take action to remove or mitigate the effect of those barriers. The next question in our framework is: what government policies could be most effective in addressing these opportunities to support the generative AI sector?

In order to answer this question, we need to consider:

- What types of policies can be used to address the barriers identified in the previous chapter?
- Are there specific options within those types of policies that are most likely to be effective in the case of generative AI? How quickly could the impact of these options materialise?
- What is the strategic positioning sought by policymakers and their appetite for risk, given the UK's current capabilities and barriers to growth in the generative AI sector?

In chapter 5, we found that there were opportunities for policymakers to support the growth of the generative AI sector in the UK by:

- Supporting investment in Research & Development and broader innovative activities;
- Supporting the AI science base;
- Supporting the development of AI skills;
- Supporting access to finance for new ventures in the AI space;
- Promoting the safety of AI products and services and supporting justified trust in AI systems;
- Ensuring that the UK's digital infrastructure is AI-ready and future-proof; and
- Promoting secure access to data to support the development of foundation models and generative AI applications.

While a full examination of each of these policy areas is beyond the scope of this report, in this chapter we provide a high level review of the policy options available within each of these areas, the available evidence on the effectiveness of those options, the likely time frame within which the impact of these policy options would be seen, and implications for supporting the growth of the generative AI sector specifically.

To support our assessment, we have undertaken a short, targeted review of evidence on the effectiveness of relevant policies, and analysed the AI policy announcements made in the last 5 years in selected major economies.[89] All of the countries included in our review of AI policy announcements have announced initiatives that pursue all or nearly all the objectives listed above. The main aim of this chapter is to provide guidance to UK policymakers on how to prioritise these different policy areas and specific options within each area, given existing evidence on the effectiveness of these policies and the UK's current AI capabilities.

6.2 Policy options and evidence on their effectiveness

A summary of our findings on policy options and evidence on their effectiveness in supporting the development of innovative economic activity is shown in Table 3 below.[90] We indicate with "high" cases where there is a substantial amount of existing evidence on the effectiveness of that policy type, and this evidence indicates relatively large effects. Conversely, we indicate with "low" cases where there is scarcer evidence and/or the evidence indicates small effects. Cases with "limited existing evidence" are those where there are even fewer directly relevant evaluations of the policy initiatives, and therefore we cannot comment on their likely effectiveness based on existing evidence. We describe this evidence in more detail in section 6.3.

Note that here we look at effectiveness in the sense of the likelihood that a lever has the intended impact, and the size of that impact. We do not examine the cost of the levers, which would need to be considered as a next step.

89 This review included the UK, Germany, Spain, France, Ireland, Italy and the USA.

90 In this report, we primarily assess effectiveness in terms of evidence of impact. In future applications of the framework it would also be useful to assess how the impact of these policies compares to their cost.

Table 3 Summary of findings on policy areas

Policy area	Evidence on effectiveness	Likely time frame for impact
Supporting private sector investment in innovation	Medium to high	Short to medium term
Funding for academic research in AI science base	Low to Medium	Medium to long term
Supporting start-up and scale-up access to finance	Medium	Short term
Skills policies: Funding STEM and AI-related advanced education	Medium	Medium to long term
Skills policies: attracting skilled workers from abroad	High	Short term
Actions to promote AI safety	Limited existing evidence	Short to medium term
Policies on access to data	Limited existing evidence	Short to medium term
Providing infrastructure	Limited existing evidence	Medium to long term

Source: Frontier Economics

In summary, our initial analysis finds that existing evidence of effectiveness (especially in the short term) is strongest for:

- Policies that aim to support private sector innovation through loans or grants, especially when provided to SMEs;
- Policies to attract high-skilled talent from abroad; and
- Supporting start-ups and scale-ups with access to finance

Support to SMEs and to start-ups and scale-ups are likely to be somewhat more relevant for the foundation and application layers, compared to the compute layer, due to the large economies of scale in compute which may make it challenging for smaller enterprises to compete.

However, it is important to note that looking at existing evidence of impacts should be only one part of the process of selecting effective policy initiatives. For example, there is limited directly relevant evidence on past policies that tells us whether promoting the safety of generative AI will be effective at fostering the growth of the sector, however, this should not necessarily deter policymakers from considering promoting safety as an important way to grow the generative AI sector in the UK. It only tells us that promoting safety has not yet been

proven as a way to support the growth of economic activity to the same extent that policies to directly support innovation have.

Overall, this initial application of our framework suggests two possible strategic approaches to prioritising different policy options.

The first approach would be to prioritise policy options that support the development of the UK's existing relative strengths. This suggests a focus on the application layer of the generative AI value chain. The UK's existing capabilities in this layer are somewhat stronger than in other layers and these capabilities could be leveraged in developing a comparative advantage in the UK for developing AI applications in the short to medium term (particular in areas such as fintech and biotech). There are also barriers to growth that are likely to have a stronger relative impact on innovative SMEs in the application layer and good evidence that government support can be particularly effective at addressing these types of barriers – for example, providing financial support for innovation at the application layer and providing transparency and assurance around the safety of AI applications.

This approach would give priority to policy initiatives that are well evidenced and build on the UK's current areas of strength. As such, these initiatives may be more likely to achieve their objectives in the short to medium term and less likely to require very high costs to be effective. The risk with this approach is that by focussing on areas of current strength, policymakers may be neglecting opportunities to broaden the UK's capabilities, and this may limit the UK's capacity to be active across all parts of the generative AI value chain in the future.

Given the dependencies identified in chapter 4, a natural complement for this approach would be to monitor the requirements of the UK's foundational and application layer as users of compute infrastructure, and act where needed to ensure those requirements are met. However, this would not extend to investing in the compute layer with the objective of fostering a UK presence in that layer as a provider of high-performance AI computing goods or developing the UK into a major provider of compute as a service to international users.

The second potential approach would be to prioritise policy initiatives in areas where the UK currently has more limited capabilities, such as the foundation and compute layers. There are potentially significant benefits from greater participation in these layers, however, UK capabilities for participation are somewhat limited and there is a potential role for government in addressing barriers to the development of these capabilities.

This alternative approach would help ensure the UK can be active across all layers of the generative AI stack in the medium to longer term by addressing key areas where the UK is further behind. In the long term, it could help ensure that AI developers in the UK can access sufficient computing power, although it may not be the most cost-effective way to achieve this objective. Moreover, this approach also hedges against the considerable uncertainty that exists around which parts of the generative AI value chain will see the most impactful developments in future and whether the UK will be able to rely upon access to foundation models and compute capacity developed in other countries (as discussed in chapter 4).

The downside of this alternative approach is that impactful policy initiatives in this area may be considerably more expensive due to the large economies of scale involved in manufacturing computing hardware and providing computing as a service, the need to be at the leading edge of technology in the design and manufacture of processing units, and the cost of purchasing processing units (likely to be very expensive due to current global chip shortages). Additionally, the benefits of this approach would only be likely to materialise in the longer term and the approach risks failing to capitalise on current areas of UK strength.

In principle, these two alternative approaches are not necessarily mutually exclusive, and indeed a combination of both is likely to be desirable. However, in a context of tight public budgets there will likely need to be some prioritisation between these two areas. To the extent that UK government AI policy to date has focused somewhat more on the development of computing and foundation model capabilities (for example, its commitment to spending £1.5 billion on expanding high-performance computing facilities and a dedicated AI Research resource), it may be that further investment and support could be most effectively targeted in the application layer, capitalising on the UK's areas of relative strength.

Below we describe in more detail the potential policy options available, evidence on effectiveness and timeframes, and implications for generative AI. We subsequently set out the specific policy options that we recommend as high priority in our conclusions in chapter 7.

6.3 More detail on policy options and evidence on their effectiveness

6.3.1 Supporting private sector investment in innovation

Policy options

Governments can support private sector innovation in a number of ways, including:[91]

1. Reducing the effective cost of carrying out innovation activities, through:
   - Direct provision of funding (e.g. grants for R&D, loans);
   - Tax credits; and
   - Providing non-financial inputs into innovative activities at no cost or a subsidised cost, for example providing access to public sector research labs or providing access to super-computing facilities.
2. Increasing the supply of research undertaken by universities and the public sector – this research can then be used by private sector firms, decreasing their cost of research, and/or its effectiveness (e.g. when academic research provides methods that can be used in private sector research);
3. Increasing the supply of skills available for private sector firms to perform innovative activities; and

[91] This taxonomy is based on Bloom, N., Van Reenen, J., & Williams, H. (2019). A toolkit of policies to promote innovation. *Journal of economic perspectives*, 33(3), 163-184.

4. Broader economic policy that increases firms' ability and incentives to innovate and reduces the cost of doing so, ranging from trade policy to intellectual property protection.

In this section, we focus on the first category, while we discuss funding for academic and public sector research (category 2) in section 6.3.2 and we discuss skills (category 3) in section 6.3.3. Much of this discussion focusses on R&D specifically because this is the type of innovation investment on which there is by far the most available evidence.

To keep the discussion manageable within the scope of this report, we do not discuss the role of broader economic policies (category 4 in the list above) in detail. However, ensuring broader UK economic policy is designed in a way that gives businesses the right incentives to invest is of course very important for the development of all sectors of the economy, including generative AI. While beyond the scope of this report, Government should also continue to consider how issues such as trade policy, merger assessment, and intellectual property protection may impact innovative sectors such as generative AI.

Evidence on effectiveness and time frames

There is a considerable body of evidence on the effectiveness of R&D tax credits and R&D grants, loans and subsidies in increasing private sector innovation activity.

R&D tax credits are a widely used policy for incentivising innovation activity, allowing businesses to write-off more than 100% of R&D expenditure against corporate tax bills. R&D tax credits have been shown by numerous studies to increase average R&D expenditure by businesses.[92] However, a concern with this finding is whether this truly reflects additional R&D activity or whether firms may simply re-label existing expenditures as R&D to benefit from tax credits. To address this, some researchers have looked at whether increased R&D spending due to tax credits also generates additional innovation outputs (such as patents, increased productivity and employment). While the available evidence is more limited, studies in this area find that tax credits do indeed also have positive impacts on innovation outputs.[93]

R&D grants, loans and subsidies are also commonly used policy tools for supporting private sector innovation, directly funding or subsidising R&D by businesses. An advantage of direct funding over tax credits is that they can be targeted to high priority industries and projects likely to have substantial economic benefits and spill-overs. However, the disadvantage of this is that it can be costly and difficult for government to correctly identify high social value projects and to monitor their delivery. Empirical evidence suggests that R&D grants, loans and

92 See What Works centre for local economic growth (2015), Innovation: R&D tax credits, page 11; and Teichgraeber and Van Reenen (2022), A policy toolkit to increase research and innovation in the European Union, page 18.

93 Innovation: R&D tax credits. (What work centre for local economic growth 2015), page 25

subsidies can be effective in increasing both R&D expenditure and innovation outputs (such as patents, productivity and employment).[94]

For both R&D tax credits and direct funding for R&D, there is evidence that these policies may be more effective in increasing innovation in small and medium size firms.[95] This may be due to the greater financial constraints faced by SMEs and the additional difficulty they can face securing capital funding, as discussed in chapter 4. It may also partly reflect that for larger firms, public support makes up a relatively small amount of overall R&D spend, so positive effects are harder to detect.

There is also evidence from a 2021 study by Ipsos MORI for UKRI on the effectiveness of R&D grants in the context of AI specifically. The findings of this study are broadly in line with the general findings described above. In particular, this study found that in the year in which a grant was received, businesses exhibited a 9% increase in R&D spending and a 13% increase in R&D employment on average.[96] There was also some evidence that impacts are larger for smaller firms, with the R&D spending and employment effects of grants only being on-going after a year for businesses with 10-49 employees and not larger companies.[97]

Implications for generative AI

In the context of the generative AI value chain, evidence of increased effectiveness of R&D tax credits and R&D grants, loans and subsidies for SMEs suggests that these policies may be more effective in supporting the application layer. This is because the application layer exhibits lower compute requirements than are needed for pre-training of foundation models and is likely to support a larger number of SMEs developing generative AI applications for different sectors. Additionally, evidence suggests that the (short-term) effectiveness of R&D grants for AI projects depends on barriers to commercialisation.[98] To the extent that barriers to commercialisation are lower in the application layer (due to the greater customer facing nature of activity in this layer), this suggests R&D grants may be more effective in the short term when targeted toward the application layer.

The size of funding required to support successful R&D projects for different parts of the generative AI stack is a further consideration. Innovate UK provided £280m in funding for 550 AI projects between 2017 and 2020, meaning the average size of a grant award was just

94 See What Works centre for local economic growth (2015), Innovation: grants, loans and subsidies page 35; and Teichgraeber and Van Reenen (2022), A policy toolkit to increase research and innovation in the European Union, page 21.

95 See What Works centre for local economic growth (2015), Innovation: R&D tax credits page 31; What Works centre for local economic growth (2015), Innovation: grants, loans and subsidies page 6; and Teichgraeber and Van Reenen (2022), A policy toolkit to increase research and innovation in the European Union, page 21.

96 Impact review of Innovate UK's AI-related activity. (Ipsos MORI 2021), page 87

97 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021), page 48

98 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021)

£500k per project.[99] If the anticipated costs of an average R&D project in different layers of the generative AI stack varies substantially, this might have implications for the relevance of this form of funding. For example, as R&D projects in the compute and foundation layer are likely to be substantially more expensive, available grants may be insufficient to make a meaningful impact to R&D in these layers. As such, R&D grants may be more effective if targeted at the application later, unless the size of individual grants is increased significantly.

In the case of tax credits, these are well established within the business tax system in a mostly sector agnostic form and there would be a number of challenges with targeting additional R&D tax relief towards generative AI businesses specifically (for example, challenges with verifying which businesses operate in this sector). However, as generative AI applications have the potential to be applied across many industries, the sector agnostic nature of tax credits may be effective in supporting businesses developing generative AI applications that are not traditionally focused on R&D and are less likely to be looking for or applying for R&D grants. Therefore, it may be most effective to maintain the current approach to tax credits and use grants and loans for more specific targeting of AI related projects.

6.3.2 Funding for public research

Policy options

Funding for public research includes ongoing general funding for research in universities, such as through the Research Excellence Framework, as well as specific grants that universities and academics can apply for. As well as universities, public research also occurs in and can be funded through national laboratories and other government bodies (such as the National Physical Laboratory and the UK's Catapult Network).

Funding for public research is intended to increase total research activity and associated knowledge spill-overs, while also increasing the availability of high-level skills in the economy. In this sub-section we focus on the first of these two effects, and we discuss promotion of AI skills further in the next sub-section.

In some cases, grant funding for public research puts a specific focus on collaboration with the private sector and commercialisation, however, generally research funded in public organisations is expected to have longer lags to commercialisation and economic impact. In the case of grants for specific projects, as with grants for private sector R&D, these have the advantage that funding can be targeted towards priority areas and research likely to have greater knowledge spill-overs. However, this creates an additional challenge for government in identifying which projects are likely to have the greatest benefits and knowledge spillovers. In the case of general funding for universities, the challenge is to ensure that the correct incentives are in place to promote high impact research. Appropriate incentives for

[99] Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021), page 15.

encouraging spin-outs and commercialisation are also important, such as giving academics rights for commercialisation of inventions and innovations funded by public research.[100]

Evidence on effectiveness and time frames

Studies looking at the impact of academic grants tend to find positive but small effects on research output.[101] For example, a study of National Institutes of Health (NIH) grants found that these lead to an approximately 7% increase in academic publications over five years.102 Quantitative evidence on economic impacts of academic funding is more limited, however, this is largely due to the longer time frame between academic research and economic impacts, making it more challenging to quantitatively identify the causal impact.

There is also evidence that funding for academic institutions can generate local impacts and spillovers, potentially leading to geographic clusters of research and innovation – a famous such example being Silicon Valley.[103] As private sector businesses locate near universities they can benefit from availability of skilled workers, research collaboration with the public sector and knowledge spillovers.

Implications for generative AI

As discussed in chapter 3, the UK has a historically strong academic science base that can be leveraged in growing the generative AI sector, however, there is some evidence of relative declines in the impact of UK academic research (both in AI specifically and more generally). As such, there may be a role for additional grants targeted at academic research in generative AI to boost innovation and research outputs in this area.

In terms of whether this support would be best targeted at certain layers of the generative AI value chain, there is no strong evidence to suggest that support for public research in any one layer would be more impactful than any other. This said, it may be beneficial to target academic research grants for generative AI towards research in sectors that are traditionally less engaged with academic research but where the UK has an established presence. For example, despite the UK's relative strength in the financial services industry, Innovate UK grant applications pertaining to AI projects are considerably higher in other areas, such as digital health technologies.[104] While this could reflect a lower need for support in early-stage development in comparison to other sectors, it may also reflect weaker linkages between the financial sector and academic research.

---

100 Hvide, and Jones (2018), University Innovation and Professor's Privilege. American Economic Review 108 (7): 1860–98.

101 Teichgraeber and Van Reenen (2022), A policy toolkit to increase research and innovation in the European Union.

102 Jacob and Lefgren (2011), The Impact of Research Grant Funding on Scientific Productivity. Journal of Public Economics 95 (9–10): 1168–77.

103 Valero and Van Reenen (2019) The Economic Impact of Universities: Evidence from Across the Globe, Economics of Education 68: 53–67.

104 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021), page 35.

It should also be noted that there is a global trend in research on foundation models increasingly being conducted within private sector organisations, rather than by universities.[105] This reflects the substantial compute requirements of pre-training foundation models and suggests that, even with simultaneous public investment aimed at increasing the access of academic institutions to compute infrastructure, if foundation models continue to grow in size and compute requirements, funding for public research may be less effective in promoting innovation within the foundation layer, and may be better targeted toward the application layer.

To the extent that government wants to use public sector research funding to create innovation clusters, it will also need to consider which geographic areas to prioritise and what factors are relevant in creating such clusters, beyond academic funding alone. To date, available evidence suggests that London has homed roughly 65% of UK AI startups since 2000.[106]

6.3.3 Improving access to finance for new ventures

Policy options

Policies to improve access to finance for new ventures can include direct grants and loans to innovative businesses, as discussed in Section 0. Such funding can often act as 'seed' investment, improving firms' ability to leverage additional funding by securing follow-on funding or capital investment.

Other policies for helping new ventures access finance include soft business support services offering advice on funding opportunities and programmes to help start-ups develop the necessary skills for investment pitches which can help secure capital/equity funding. For example, UKRI currently runs a program called Innovate UK EDGE which provides this form of business support for innovative SMEs, including running 'Pitchfest' events to assists firms in preparing for investment pitches.[107]

Evidence on effectiveness and time frames

There is relatively strong evidence that direct grant support is effective in helping firms secure private finance. In the context of AI specifically, econometric analyses using PitchBook data on equity investments secured by firms between 2008 and 2018 suggested that Innovate UK grants to businesses for AI projects increased the total equity investment raised by these businesses by 5.3 to 16.4 percent on average.[108]

105 Benaich and Hogarth (2022), State of AI Report.

106 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021), page 27.

107 https://www.innovateukedge.ukri.org/Funding-and-finance-taking-strategic-approach/Innovate-UK-EDGE-Pitchfest-Get-investment-ready

108 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021).

Research into the effectiveness of soft forms of business support and access to finance training is more limited at present.

Implications for generative AI

As discussed in chapter 5, barriers to access to finance for new ventures are likely to be somewhat more important for the application layer of the generative AI value chain. A mix of direct financial support and soft business support could play a helpful role in alleviating these barriers and supporting the growth of the generative AI sector in the UK. This said, as discussed in chapter 3, currently the UK performs relatively well in terms of the availability of venture capital investment. As such, improving access to finance may not need to be a priority area for government support in the short term, although it would be worth considering initiatives to broaden the areas of focus of VC investment.

6.3.4 Promoting AI skills

Policy options

As discussed in chapter 3, there is some evidence that UK workers have relatively good AI skills compared to other countries, however, there is also evidence that many businesses are experiencing gaps in the availability of AI skills. As described in chapter 5, there may be underinvestment in skills for various reasons, including: employers and workers not being able to predict which skills will be valuable in future; employers facing a risk of workers they train moving to other businesses; and employees not being able to credibly convey or demonstrate their skills and abilities to employers.

Policies to promote AI skills and address these barriers could include:

- organising and funding AI specific technical training programs;
- policies to incentivise and increase the number of university graduates in STEM subjects;
- funding for an increased number of post-graduate students specialising in computer science and AI;
- policies aimed at attracting more highly skilled labour from abroad (for example, making it easier for businesses to secure visa's for the family of high skill workers);
- facilitating the creation of specific qualifications in AI skills, helping workers to credibly demonstrate their abilities;
- additional government backed loans for workers to upskill and re-train in AI skills; and
- policies aimed at highlighting the demand for AI skills and encouraging colleges and universities to incorporate AI skills within a wide variety of subjects and courses.

The UK has already invested a significant amount of funding towards AI education and skills. As part of the 2018 Industrial Strategy AI Sector deal, £110 million of government funding was put towards funding 16 new AI Centres for Doctoral Training and delivering 1,000 new PhDs

over 5 years.[109] The Alan Turing Institute has also announced £46m to develop the next generation of top AI talent as part of the UK AI Sector deal.[110]

Evidence on effectiveness and time frames

A full review of empirical evidence on the effectiveness of skills policies is beyond the scope of this report, however, empirical evidence suggests that increasing the number of STEM graduates increases labour market outcomes and innovation, but the effect can take a long time to materialise. For example, a study of colleges in Norway found that the foundation of new STEM-focused colleges increased R&D and STEM related technological progress, but these effects only materialised after around 10 years.[111]

There is also considerable evidence that immigration is particularly important to increasing the innovative capacity of economies. In the US, immigrants account for 14% of the workforce but 52% of STEM doctorates, a quarter of all patents and a third of all US Nobel Prizes - and this simple observation is reflected in numerous statistical studies that find evidence of a statistically significant impact of immigration (especially high skill immigration) on innovation activity.[112] What is more, by focusing on attracting skills from abroad, immigration policies can act to increase the availability of high level skills more quickly than policies aimed at education.

Implications for generative AI

Availability of AI skills is important for all layers of the generative AI value chain and a combination of policies aimed at boosting AI skills, including shorter-term policies focused on immigration policy and longer-term policies focused on education, could be effective in supporting growth of the generative AI sector.

There is no clear evidence to expect that such policies would be more effective at supporting one layer of the generative AI stack over any other. That said, as discussed in chapter 4, the skills relevant to the development of the application layer are somewhat broader and more uncertain than in other layers. As such, it may be important to fund policies with a broader focus than just the traditional AI focus areas of computer science, data science, and mathematics. Going forward, it will be important to continue to monitor and develop new evidence on where in the generative AI value chain skills gaps are largest and what skills are most in demand.

While not a focus of this report, it should also be noted that policies aimed at increasing the general level of basic IT skills across the economy could have a significant impact on adoption

109 https://www.gov.uk/government/news/next-generation-of-artificial-intelligence-talent-to-be-trained-at-uk-universities#:~:text=1%2C000%20students%20will%20have%20the,CDTs%20)%2C%20located%20across%20the%20country

110 https://www.gov.uk/government/publications/turing-artificial-intelligence-fellowships/turing-artificial-intelligence-fellowships

111 Carneiro, Liu, and Salvanes (2018), The Supply of Skill and Endogenous Technical Change: Evidence from a College Expansion Reform. Discussion Paper Series in Economics 16/2018, Norwegian School of Economics.

112 Teichgraeber and Van Reenen (2022), A policy toolkit to increase research and innovation in the European Union.

of generative AI technologies by businesses. This would generate derived demand in the generative AI value chain, supporting its growth, while also providing potentially substantial economic benefits (see section 2.6 above). As such, government's AI skills strategy should also consider broader policies for enhancing basic IT skills, for example at the secondary education level.

6.3.5 Investment in computing and connectivity infrastructure

Policy options

As discussed in chapter 4, compute capacity and connectivity infrastructure are crucial inputs for the foundation and application layers of the generative AI value chain, both for the training and pre-training of models and also for the deployment of generative AI applications. The UK government is already investing considerable amounts in improving UK digital infrastructure, including the establishment of the Hartree National Centre for Digital Innovation (HNCDI), which represents a £210 million initiative comprised of public (UKRI) and private funding (IBM) (started in 2021 and spanning 5 years).[113] More recently the UK government has announced plans to invest £1.5 billion toward developing the UK's high performance computing capacity, much of which is directed towards a new super computer based in Bristol to drive AI innovation.[114]

Beyond these direct investments in computing infrastructure, government could also consider policies to unlock potential barriers to growth of the UK data centre sector, such as looking at reforms to planning regulations around data centres and ensuring adequate/dedicated renewable energy provision for data centres (something that is particularly important for the largest data centre providers, which have all made significant commitments to the sustainability of their data centres).

Evidence on effectiveness and time frames

In terms of evidence on the productivity impact of connectivity and compute infrastructure, there is good evidence to show that increased access to high-speed broadband (i.e. connectivity) and use of cloud services (either for data storage or compute purposes) by businesses can increase productivity and economic growth. For example, an OECD study has shown that a 10 percentage point increase in adoption of high-speed broadband (or cloud computing) in a country is associated with a 5.8 percent (or 3.5 percent) higher productivity level for the average firm after 5 years.[115] However, there is more limited evidence on the likely

[113] https://www.ukri.org/news/new-210-million-centre-to-advance-ai-and-quantum-computing/

[114] https://www.gov.uk/government/news/bristol-set-to-host-uks-most-powerful-supercomputer-to-turbocharge-ai-innovation#:~:text=Both%20Isambard%203%20and%20Isambard,%2C%20Bristol%2C%20Cardiff%20and%20Exeter.

[115] Gal, P., et al. (2019), Digitalisation and productivity: In search of the holy grail – Firm-level empirical evidence from EU countries, OECD Economics Department Working Papers, No. 1533, OECD.

effectiveness of policies aimed at attracting additional investment in data centres and connectivity.

The likely impact of direct investments by government in super-computing infrastructure is also uncertain. Due to global chip shortages and the high price of high-end GPUs, these investments are likely to be very expensive. Additionally, as mentioned above, even with these investments, if the size of cutting-edge foundation models continues to grow, it is not clear that public sector research in foundation models will be able to compete with private sector research.

Implications for generative AI

Given the considerable existing UK commitments to developing public sector super-computing capacity, the likely high cost of these investments, and the uncertainty around the impact they will have on innovation in the foundation layer, further government support for digital infrastructure may be better targeted at unlocking barriers to growth in the data centre sector and incentivising private sector investment in compute. This could help grow the compute layer of the generative AI value chain in the UK (in terms of the provision of 'compute as a service'), while also ensuring that there is adequate data centre capacity for the fine-tuning and deployment of generative AI applications.

6.3.6 Promoting AI safety

Policy options

A full examination of potential AI safety risks and policies to address AI safety are beyond the scope of this report, however, as discussed in chapter 5, addressing AI safety risks will be a crucial part of supporting growth in the generative AI sector.

At a high level, policy initiatives in this space include the design of regulations or voluntary standards, which could either relate directly to the development of AI models and applications, or relate to the application and use of AI in specific sectors. Policies could also include the development of regulatory sandboxes, as envisaged in the Government's AI White Paper.[116]

Furthermore, ensuring the safety of generative AI systems is likely to involve a broader AI safety ecosystem beyond what one might think of as the "core" generative AI sector (computing, foundation and application layer as described above). This broader ecosystem may include, for example, development of software that complements generative AI applications, cybersecurity services (research, consulting, others), management consulting and legal services. The Centre for Data Ethics (CDEI), for example, considers the development of an AI assurance market as a priority area for ensuring the safe and trusted

[116] AI regulation: a pro-innovation approach, DSIT, March 2023.

deployment of AI in the UK.[117] As such, policies that can support the growth of an AI assurance sector are also likely to be crucial in addressing AI safety and security.

Evidence on effectiveness and time frames

Evidence is currently limited on the effectiveness of different policy options relating to AI safety, however, there is reasonable evidence to suggest that the impact of addressing AI safety risks and engendering trust in the AI ecosystem could be substantial. For example, a Frontier Economics report for the Open Data Institute conducted a meta-analysis of the impact of trust on willingness to share data, finding that higher levels of trust are strongly correlated with willingness to share data and the economic value of data ecosystems.[118]

There is also reason to believe that the direct economic value of an AI assurance sector could be substantial. For example, by way of comparison, in 2021 there were 30,000 people in the UK working in the data assurance ecosystem, and these jobs are typically high-productivity jobs (measured by output or Gross Value Added per job) in the information technology sector.[119]

Implications for generative AI

As discussed in chapter 5, addressing AI safety risks is important to all layers of the generative AI value chain but there are several reasons why these issues are likely to be particularly important (especially in the short term) to the application layer. Clear regulation or voluntary standards around data protection and safety could alleviate trade barriers and increase demand for generative AI applications, helping to grow the generative AI sector in the UK.

Additionally, to the extent that the UK may be especially well-placed for the development of generative AI applications in financial, professional and legal services, as well as health and biotechnology, UK policymakers may want to consider with high priority how to promote AI safety in these areas. From an economic point of view, this may be a higher priority for the UK than for other countries that may specialise in other areas of application for generative AI. This could include, for example, prioritising health and/or financial applications of generative AI as a focus of regulatory sandboxes.

6.3.7 Access to data

Policy options

As discussed in chapter 5, barriers to data access and policies affecting access to data are also important factors in the growth of the generative AI sector. While a full examination of policies affecting data access is beyond the scope of this report, relevant policies include:

117 Centre for Data Ethics and Innovation (2021), The roadmap to an effective AI assurance ecosystem.

118 Frontier Economics for the Open Data Institute (2021), The economic impact of trust in data ecosystems.

119 Frontier Economics for the Open Data Institute (2021), The economic impact of trust in data ecosystems.

openness of access to government data sets, IP and copyright law, and programmes to support business in sharing valuable data securely and ethically (potentially including regulatory sandboxes).

A very common way for AI business to secure data access is through direct collaboration with business or organisations that hold valuable data. Government policies can look to support such collaboration by organising programmes such as DataPitch, an EU-funded accelerator programme which provides start-ups with ideas for data-driven products together with support including initial funding and matching to potential data providers. Direct government funding for R&D could also be effective in facilitating such collaboration.

Currently the UK is second only to Canada on the Open Data Barometer, a global measure of how governments are publishing and using open data.[120] However, qualitative research interviewing recipients of UKRI funding for AI projects found that stakeholders consistently felt more work needs to be done to simplify access to relevant data and access to data was the most commonly cited barrier to commercialisation.[121]

Evidence on effectiveness and time frames

Evidence on the effectiveness of policies to improve access to data is currently limited. There is some evidence that regulatory sandboxes can be effective, with the Financial Conduct Authority's Regulatory Sandbox being frequently cited as a good example of how real time data could be made available more readily to developers in a safe and secure setting, without impacting financial markets. [122]

There is also evidence that the EU funded accelerator programme DataPitch, mentioned above, has had positive economic impacts.[123] Analysis of UKRI funding for AI projects found that half of the firms supported were "*motivated to collaborate*" in order to obtain the necessary data and noted that the funding was key to progress.[124]

Implications for generative AI

Data availability is crucial as an input for both the foundation and application layers of the generative AI value chain and there is a clear role for government in helping facilitate access to high quality data in secure and ethical manner. In the shorter term, facilitating collaboration and access to sector specific data for fine-tuning and applications of generative AI is likely to be particularly important. In the longer term, as foundation models use up more and more of

120 Open Data Barometer – 4th edition.

121 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021).

122 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021).

123 London Economics (2020), "Evaluation of Data Pitch".

124 Impact review of Innovate UK's AI related activity. (Ipsos MORI 2021).

the high quality publicly available data for pre-training, proprietary data may become increasingly important to the foundation layer as well.

Evidence gaps and uncertainties

In this chapter we have addressed the fourth and final factor of our framework, looking at what policy interventions could be most effective in alleviating the barriers to growth in the generative AI sector. As highlighted throughout this chapter here are considerable evidence gaps relating to the likely effectiveness of different policy options, particularly policies relating to AI safety, provision of compute infrastructure, and access to data. Where possible, further work should look to assess the effectiveness of past policies in these areas (to the extent there is relevant policy precedent) and also build evaluation into the design of new policies in these areas, and other policy areas discussed in this chapter.

7 Conclusions

Our report has provided an initial application of an economic framework designed to help policymakers take decisions that would maximise the economic benefits of generative AI to the UK. We also provide recommendations on the types of evidence Government may wish to monitor to update its thinking over time.

7.1 Initial recommendations on prioritisation

The UK has significant capabilities that could enable the growth of the generative AI sector, but the relevance of these capabilities does not fall equally across the layers of the generative AI value chain. Most notably, the UK's existing capabilities are likely to lend themselves most directly to the development and deployment of generative AI applications, especially in the financial, professional services and health sectors where the UK has an existing comparative advantage that could be built on. In contrast, the UK has relatively limited capabilities in the manufacturing and assembly of computing hardware, and in the provision of high-performance computing as a service. To date, UK presence in the development of foundation models has been limited, but as the UK has a strong science and skills based in AI by international standards, it may be possible to develop a stronger presence in this space.

There are clear areas where government policy could support the development of generative AI applications to ensure that the UK capitalises on its potential strengths in this layer. However, there is a broader question about whether the UK should also be considering taking action to build up strength in the areas that may support the development and evolution of the foundation and compute layers. The answer to this question depends on when government would want its initiatives to have an impact (short term, i.e. around 5 years from now, or longer term), on its appetite for risk (how it prioritises actions that could have either a large positive impact or no impact, compared to actions with less upside but more certain returns) and on strategic fit against broader policy and government objectives.

Our assessment is that a combination of both actions to capitalise on existing strengths in the development of generative AI applications, and to support the development of foundation and compute layers is likely to be desirable. However, it is important to recognise that actions aimed at the compute and foundation layers are likely to take longer to bear fruit and are subject to greater risk.

Therefore, we recommend assigning higher priority to actions directly supporting the application layer, while also monitoring the requirements of application developers as users of compute infrastructure and foundation models and acting if needed to ensure those requirements are met. In particular, our preliminary view of the evidence is that high priority actions for government should include the following:

- Direct support to business innovation in the development of generative AI applications, building on existing policy tools that have been shown to be effective, such as grants for R&D;

- Initiatives to promote AI safety in key areas of application for the UK, such as financial services and health. This could include for example prioritising health and/or financial applications of generative AI as a focus of the regulatory sandbox(es) envisaged in the Government's AI White Paper ("A pro-innovation approach to regulating AI").

- Investing in training and attracting top-level talent, especially in key skill areas for application development, such as data engineering and prompt engineering.

- Ensuring that developers of generative AI applications can access data, foundation model and compute resources at the level required. Pursuing this objective could include:

    - Monitoring the likely compute demand of the AI applications sector and taking action if needed to address bottlenecks that may prevent or delay supply from meeting this demand.
    - Monitoring and, if necessary, taking actions to address the availability of foundation models of the required performance level and required level of access for developers of AI applications (whether this is access that enables fine-tuning of the model weights or more limited access).
    - Monitoring the extent to which access to data that is currently not collected or not available to AI application developers would accelerate the development of the sector (e.g. for model fine-tuning) and exploring opportunities for facilitating access to such data.

Actions aimed at developing the UK's participation in the foundation layer (beyond those already announced, such as the government's £1.5 bn investment in compute capacity and the creation of the AI Safety Institute) should be considered lower priority, however, could still be worth pursuing. Such actions include:

- Funding research specifically on foundation models and transformer architecture that could feed into the next generation of foundation model development.
- Further investment (beyond what has already been announced) in directly providing researcher access to powerful supercomputers.
- Prioritising support to private sector initiatives to build compute capacity for pre-training of foundation models specifically (rather than for AI inference or other uses).
- Investing in creating large high-quality datasets for foundation model training (on top of investment in more specific datasets for fine-tuning and application development, described above).

7.2 Recommendations for further evidence gathering

Because AI is a fast-moving field, defining and implementing the best possible public policies to maximise the benefits of the sector will be an ongoing exercise rather than a task that can be accomplished fully now. Therefore, we would recommend that government continues to apply and update the framework provided in this report over time. To aid in this, we have set out a number of key questions and issues that future applications of this framework should seek to address. This includes the following:

- Developing a more precise understanding of the inputs and capabilities relevant to each layer of the generative AI value chain. For example, what specific skills are most important for each layer currently, and what skills are likely to be most important in the future?
- Developing more precise measures of the capabilities assessed in this report, possibly through new primary evidence gathering, such as surveys.
- Developing direct measures of the availability of data to AI developers in the UK.
- Assessing the current and likely future demand for and supply of UK based computing capacity (rather than global cloud computing) to help understand whether a lack of domestic capacity may prove to be a bottleneck in the future, and understand what government could do to assess any possible constraints.
- Monitoring emerging and likely sources of future progress in generative AI in order to better direct government efforts.
- Monitoring the role of open-source foundation models in both AI research and commercial application development and whether open-source models continue to provide efficient access to foundation models, including for fine-tuning to specific applications.
- Collecting further evidence on the relative importance of different barriers to the development of generative AI in the UK, possibly through surveys with a particular focus on businesses developing AI and AI applications.
- Developing additional evidence on the effectiveness of policies aimed at supporting AI (especially where current evidence is limited). In particular, prioritising evaluations of policies targeted at: i) promoting the safety of AI products, ii) developing the UK's digital infrastructure and iii) promoting secure access to data.