CMA
Competition & Markets Authority

AI Foundation Models: Initial Report

18 September 2023

© Crown copyright 2022

You may reuse this information (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence.

To view this licence, visit www.nationalarchives.gov.uk/doc/open-government-licence/ or write to the Information Policy Team, The National Archives, Kew, London TW94DU, or email:psi@nationalarchives.gsi.gov.uk.

1. Introduction....5
How we conducted the review....6
The structure of this report....6
2. Background....8
What are FMs?....8
How are FMs developed?....10
Data preprocessing and architecture design....10
Pre-training....11
Fine-tuning....11
Computing power....12
Deployment, routes to market and monetisation strategies....14
AI supply chains and vertical relationships....16
How FMs are evaluated....18
The FM landscape....20
Potential application of FMs....25
3. Competition and barriers to entry in the development of FMs....27
Introduction....27
Data requirements....28
Pre-training....28
Fine-tuning....30
Alignment....30
Domain- or task-specific fine-tuning....32
Synthetic data....32
Computational resources....33
Pre-training....34
Fine-tuning....36
Inference....37
Technical expertise....38
Access to funding....39
Open-source models....40
Pre-training....40
Fine-tuning....41
Uncertainties....41
Will access to proprietary data become necessary to compete?....42
Will models become larger?....44
Will FMs be highly generalised?....46
Is cutting edge performance required to compete?....47
Will large technology companies and first movers have an advantage over others?....48
Will open-source models remain a key part of the market?....50
Conclusion....52
4. The impact of FMs on competition in other markets....54
Introduction....54
Deploying FMs in downstream markets....55

FMs could become an important input in a wide range of markets 55
Firms can access FMs in a number of ways 55
Firms are monetising FM services in different ways 57
Potential impact of FMs on competition in downstream markets, including potential risks from vertical integration 58
How FMs could drive competition and disrupt incumbent firms 58
Different types of vertical integration and partnerships 66
Features of downstream markets that could affect competition in upstream FM development. 67
Uncertainties 70
Will downstream firms continue to have access to a wide range of FM deployment options and find it easy to switch between them? 71
Will consumers be able to make choices between FM services effectively? 72
Will consumers prefer integrated and customised FM services? 73
Would vertically integrated firms and partnerships have an incentive to foreclose upstream and downstream competitors? 74
How significant are data feedback effects in downstream markets? 75
Conclusion 77
5. Consumer Protection 79
Introduction 79
Consumer protection concerns identified in our review 79
Exacerbating existing consumer harms 80
False and misleading outputs from FMs 81
Why hallucinations happen and some examples of what they look like 82
Potential for user manipulation 84
FMs and advertising 85
Consumer understanding 86
Disclosure when interacting with a FM-generated response 88
Disclosures on the limitations of FMs 89
Technical measures to address possible consumer harms 90
Testing, evaluation and mitigation 90
Mitigating hallucinations 92
Watermarking 93
Approaches to AI governance 94
Uncertainties 95
Will consumers be able to identify if false and misleading information is provided by a FM application? 96
Will consumers know they are interacting with an FM-generated output and fully understand the risks of doing so? 98
Will new and / or existing technical solutions reduce the prevalence of false and misleading information and if so, how substantially? 99
Will consumers have clear routes to redress if things go wrong? 99
Will there be accepted standards or benchmarks to measure the quality and / or reliability of FM-generated outputs? 100

Conclusion....101
6. Competition and consumer protection law, and the role for regulation....103
Introduction....103
The legal framework in the UK and its application to AI....103
Competition law....104
Consumer protection law....105
Forthcoming powers for the CMA to better enforce competition and consumer law and increase competition in digital markets....107
Digital Regulation Cooperation Forum (DRCF)....107
The UK's policy approach to AI....109
A pro-innovation approach to the regulation of AI....110
Global summit on AI safety....112
Inquiries into AI....112
International approaches to the regulation of AI....113
Industry initiatives....116
Interaction between competition, consumer and other policy objectives....117
AI safety and competition....117
Intellectual property and competition policy....118
The ongoing role of regulation....119
7. Principles....120
Principles to guide the development and deployment of FM markets....120
8. Next steps....122
9. Glossary....123

1. Introduction

1.1 The CMA is an independent non-ministerial UK Government department and is the UK's principal competition and consumer protection authority. We help people, businesses and the UK economy by promoting competitive markets and tackling unfair behaviour.

1.2 In its 2022 Autumn Statement the Government committed to introducing the Digital Markets, Competition and Consumers Bill which will provide the CMA with powers to operate a statutory pro-competition regime for digital markets through its Digital Markets Unit (DMU).[1] The DMU, which currently supports our work across digital markets using the CMA's existing powers, has already begun work to operationalise the new regime.

1.3 This initial review fits with our proposed medium-term priorities and areas of focus, set out in our draft Annual Plan 2023/24, to:

- enable open access to markets for innovating businesses;
- help emergent sectors to develop into high growth, innovative and competitive markets; and
- prioritise sectors that offer the biggest potential for improvement in innovation and productivity.

1.4 We have been engaging with the government on its various workstreams in this area and will continue to do so, both individually as the CMA fulfills its functions and jointly with our fellow regulators in the Digital Regulation Cooperation Forum. AI and AI regulation is an active policy area of the government, and there have been several notable recent developments, including:

- in March 2023, the Government published a white paper setting out its 'pro-innovation approach to AI regulation', which called on regulators, including the CMA, to implement five principles to guide and inform the responsible development and use of AI;
- Sir Patrick Vallance, the Government Chief Scientific Adviser, published his report on 'Pro-innovation Regulation of Technologies Review – Digital Technologies', which made a number of recommendations in relation to Generative AI; and
- in the Spring Budget 2023, the Chancellor of the Exchequer referred to an announcement that the government will establish a new government-industry

[1] AUTUMN STATEMENT 2022 (publishing.service.gov.uk)

taskforce to advance UK sovereign capability in foundation models, including large language models, that has £100 million in initial start-up funding committed.

- In June 2023, the Prime Minister announced that the UK will host the first global summit on AI safety in Autumn 2023.[2]

1.5 This initial review by the CMA looking at competition and consumer protection issues in relation to AI Foundation Models forms part of our response to these developments, to build our understanding and evidence base, and help us prepare to meet the expectation from government that regulators, including the CMA, play their part in supporting innovation in AI that benefits consumers, businesses, and the UK.

1.6 The development of AI has raised several other important issues, including safety; security; privacy; intellectual property and copyright; and human rights. These issues are being considered by other regulators and Government. This review focused on questions that the CMA is mandated and best placed to address, namely questions around competition and consumer protection.

How we conducted the review

1.7 As part of our initial review, we engaged with over 70 stakeholders, including a range of Foundation Model ('**FM**')[3] developers, businesses deploying FMs, consumer and industry organisations and academics. We gathered information directly from stakeholders as well as considering publicly available information, including the latest AI research.

1.8 We would like to thank all stakeholders for their engagement with this initial review.

The structure of this report

1.9 This report sets out the technical detail of how FMs work, what is required to develop them and how they can be used in a range of products and services. The report goes on to consider the likely competition and consumer protection issues that could arise from the use of AI under three themes:

- **Theme one** – competition in the development of FMs;
- **Theme two** – the impact of FMs on competition in other markets; and
- **Theme three –** consumer protection.

[2] UK to host first global summit on Artificial Intelligence - GOV.UK (www.gov.uk)
[3] See chapter 2 for an explanation of what FMs are.

1.10 For each theme we set out a broad spectrum of possible market outcomes – ranging from the more positive to the more concerning – and identify factors which might contribute to those different outcomes.

1.11 The report also sets out current approaches to regulating AI in the UK and internationally.

1.12 It concludes by setting out clear principles that can help the market develop in an open and competitive way and inform future CMA work in this area.

1.13 Finally, we outline our next steps for further work in this area.

2. Background

2.1 This chapter will introduce FMs, including:

- what they are;
- how they are developed, including how they are trained, deployed, and evaluated; and
- the FM landscape, considering the range of models developed by a variety of companies, the performance of a selection of models against benchmarks, and a snapshot of investment into startups developing FMs.

What are FMs?

2.2 FMs are a type of AI technology that are trained on vast amounts of data that can be adapted to a wide range of tasks and operations.[4] Products and services that utilise the technology are already being developed by new and existing businesses.

2.3 Most FMs are currently being developed using a deep learning model called a transformer, first introduced by Google in a white paper in 2017.[5] However, as techniques evolve, new and improved algorithms or architectures for developing FMs may be discovered.

2.4 The type of data that is used to train a FM determines its 'mode'. For example, large language models (LLMs) are a type of FM trained on text data, and image generation models are trained on image data (coupled with text). A multi-modal FM is a FM that is trained using multiple types of data (see Figure 1). Expanding the modality of FMs to include other data such as 3D environments, video and audio is an area of ongoing research.

[4] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
[5] Vaswani, A, Shazeer, N, et al (2017), Attention is All You Need

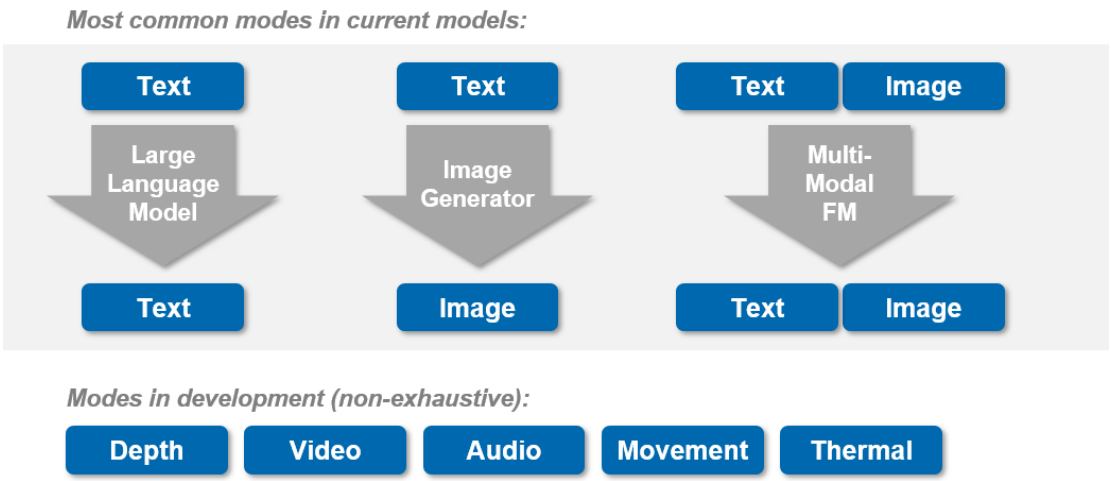


Figure 1: An illustration of the different modalities of FMs

2.5 In preparation for the training process, the vast training data sets are broken down into billions of small tokens. In the case of text data, each token may represent a word or parts of a word. During training, the model learns the probabilistic relationships between each token and every other token in the data set they are provided.[6] Using a mechanism called self-attention,[7] the model learns which tokens provide context about the meaning of others. For example, as shown in Figure 2, in the sentence 'I swam across the river to get to the other bank', the model can identify that the words 'swam' and 'river' provide context to the meaning of the word 'bank' that indicates that this instance of 'bank' does not relate to any other instances of the word 'bank' that are used in a financial context.

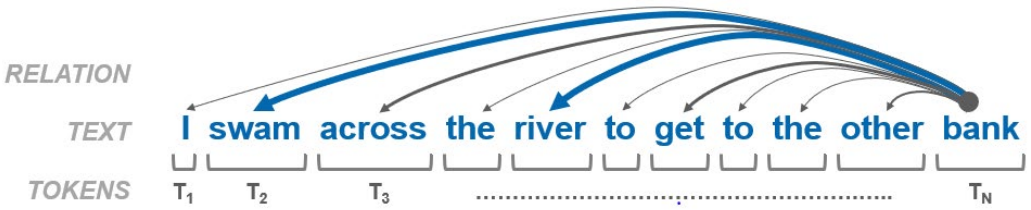


Figure 2: An illustration of the concept of self-attention, in which a thicker arrow indicates a stronger relation identified by a (hypothetical) transformer.

2.6 FMs are therefore a network of fixed calculations that convert inputs to outputs. To enable a trained model to get the correct outputs from inputs, there are individual multipliers and additions that apply to each of the fixed calculations. These multipliers (known as weights) and additions (known as biases) are a set of parameters that are iteratively adjusted during training based on the inputs and outputs provided in the training data. The number of parameters (weights and biases) in a model is the amount of information required to 'store' the knowledge of

[6] Depending on the specific architecture used, the model may only learn the probabilistic relationships of tokens that have come before it in the sequence of data.
[7] Vaswani, A, Shazeer, N, et al (2017), Attention is All You Need

the model and is therefore also referred to as the size of the model. In a small model, the knowledge is encoded in fewer parameters than in a large model.

How are FMs developed?

2.7 There are a number of steps required to develop, train and deploy a FM, illustrated in Figure 3.

2.8 For the purposes of this report, we distinguish between 'upstream FM development and supply' and 'downstream FM services'. We define the former as the level in the supply chain at which FM developers produce and distribute FMs, and the latter as the markets[8] in which FMs are deployed (in which FM developers may also compete).

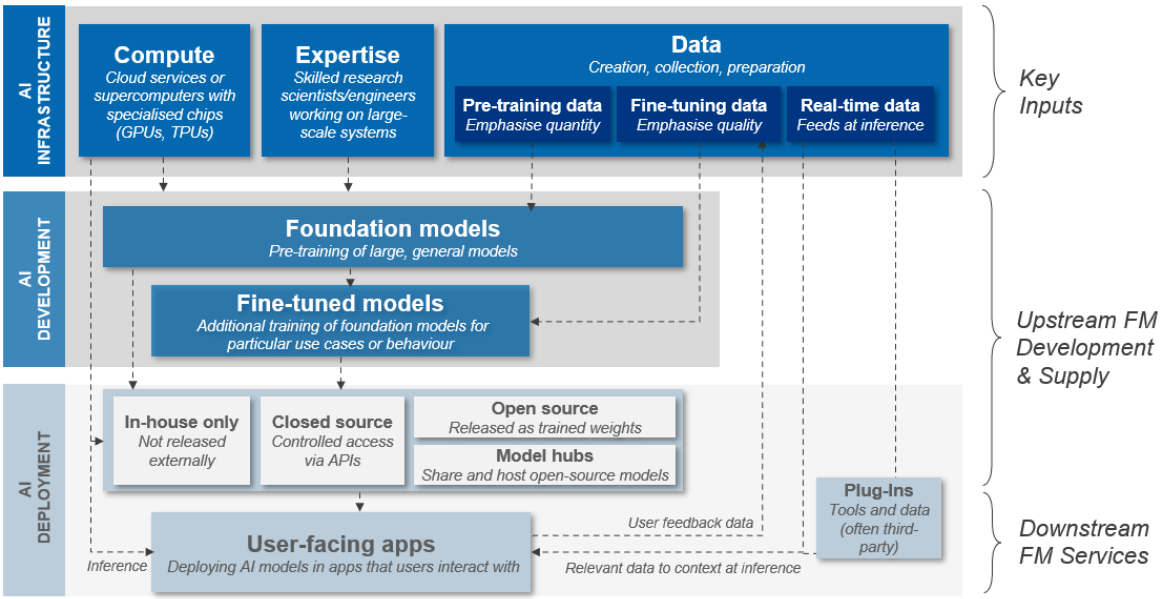


Figure 3: An overview of foundation model development, training and deployment.

Data preprocessing and architecture design

2.9 The first steps to building a FM are to design and implement its architecture, for example by deciding its size (how many parameters it will have) and its topology (the structure of the network). Engineers must also prepare the training data, by collating it from different sources and in some cases removing parts of the data that are harmful or not useful. The data is then tokenised, converting it into a format that can be used for training.

[8] We have not carried out a market definition exercise as part of this review, therefore any reference to 'markets' refers to markets in its plain English meaning rather than a defined market under competition law.

Pre-training

2.10 The first stage of training a model is called pre-training. At this stage, hundreds or thousands of gigabytes of data are used to build the knowledge of the model. Commonly, the data used at this stage is from publicly available sources, such as web crawling or open datasets, although proprietary data can also be used. In the context of pre-training, web crawling is the use of automated bots to crawl the web for new or updated webpages which are then scraped for training data.

2.11 Some notable publicly available data sets commonly used for training LLMs and image generation models include:

(a) **C4**: An English-language dataset (approximately 750 GB) prepared by AllenAI through cleaning a broader, open dataset called Common Crawl that has been built through 12 years of web scraping.[9]

(b) **The Pile:** A combination of 22 high quality datasets, compiled by EleutherAI. Sources include PubMed, ArXiv, GitHub, YoutubeSubtitles, DM Mathematics and Stack Exchange.[10]

(c) **Project Gutenberg Corpus:** A compilation of over 50,000 books in the public domain, introduced by Gerlach et al. in 'A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics'.[11]

(d) **LAION-400M:** A dataset, derived from Common Crawl, of 400M image and descriptive text pairs, collated by non-profit organisation LAION.

(e) **LAION5B**: A dataset consisting of 5.85 billion filtered image-text pairs, as well as image labels to use when training models to detect inappropriate or toxic content, collated by LAION.

Fine-tuning

2.12 Fine-tuning is an optional additional process that can be applied to pre-trained models to add specific capabilities or improvements using particular datasets. There are two main types of fine-tuning:

(a) **Alignment** is the process of fine-tuning to improve the behaviour of a model to align with the expectations or preferences that a human user may have. The types of expectations for which fine-tuning is conducted include:

---

[9] The C4 data set is available at: https://huggingface.co/datasets/c4
[10] The Pile data set is available at: https://pile.eleuther.ai/
[11] Gerlach, M, Font-Clos, F (2018), A standardized Project Guttenberg corpus for statistical analysis of natural language and quantitative linguistics.

(i) *Prevention of biased, false,*[12] *or harmful outputs*: the vast data sets used for pre-training models often contain harmful, biased or false content. In addition to filtering training data, a common technique to improve behaviour is reinforcement learning from human feedback (RLHF), which trains the model with a reward function that punishes 'bad behaviour'. This relies on human feedback to distinguish between good and bad behaviour, which can be provided by paid contractors or directly from users.

(ii) *Machine-like or conversational responses*: data used for pre-training, such as text scraped from the web, does not usually contain the examples required to teach the model to 'speak like a machine' so as not to mislead users. Therefore, examples of human-machine conversations can be used to fine-tune a pre-trained model to add this capability. These datasets can be human generated, or examples of conversations from existing chatbots or large language models can be collated and used to train new models.

(b) **Domain or task specific** fine-tuning is the process of specialising a pre-trained model to a particular domain or task. This process requires smaller, more highly curated datasets than for pre-training. For example, a data set containing legal documents could be used to improve the ability of a model to provide legal advice or generate legal documents.

Computing power

2.13 FMs use a large number of mathematical operations to calculate the output of the model, during both training and inference. In most cases, due to the size of the models and the amount of training data required, it is not feasible to train and run FMs on conventional computer chips (such as central processing units (CPUs)).

2.14 Accelerator chips can run the operations in parallel and are therefore used to speed up the computation of the FM. Conventionally, AI developers use graphical processing units (GPUs), a type of accelerator chip designed to process images efficiently. The architecture of a GPU is suited to processing deep learning models in parallel, making them more efficient to use than CPUs (at least two times faster).[13] Multiple GPUs can be used in parallel to increase the efficiency of computations, whereas CPUs do not scale well with deep learning.

2.15 NVIDIA is currently the main supplier of GPUs that are used for AI purposes. One of the likely reasons for its lead is its legacy of supporting deep learning software

[12] For non-creative applications, there is a desire to ground FM output in real-world knowledge or the source context.
[13] Microsoft Azure (2018) GPUs vs CPUs for deployment of deep learning models

packages (PyTorch, Tensorflow) to work well with its GPUs. NVIDIA's H100 GPU is its latest flagship AI accelerator, first available for sale in 2022 and started shipping in the first quarter of 2023.[14]

2.16 A range of different types of accelerator chips for AI are also being developed and used.[15] For example, Google manufactures its own hardware accelerator, the tensor processing unit (TPU). These chips are designed for deep learning applications and benefit for performance for some types of neural networks. Google considers its TPUs of similar performance with NVIDIA GPUs e.g. TPU v4 with NVIDIA's previous flagship A100 GPU.[16]

2.17 The pre-training of a single FM requires the use of a large number of accelerators (typically in the 100s or 1000s) over a number of days. Figure 4 illustrates the scale of some selected foundation models for which this type of data is available.

Model	**Parameters**	**Training data (in tokens)**	**Training time (in days)**	**Hardware (GPUs/TPUs)**
LLaMA[17] (Meta)	65B	1400B	21	2048 A100 GPU
LaMDA[18] (Google)	137B	2810B	57.5	1024 TPU v3
GPT-3[19] (OpenAI)	175B	300B	34 [estimated]	1024 A100 GPU [estimated]
MT-NLG (Microsoft/ NVIDIA)[20]	530B	270B	90	4480 A100 GPU

Figure 4 Training data, training time and hardware used for training for a selection of models of various sizes.

2.18 The compute cost for developing a given model depends on a number of variables including the model and data size, the hardware used and the selected cloud provider. Costs for compute vary across cloud providers, but also, organisations that require large commitments are typically able to negotiate compute deals at less than the on-demand rate. Given this, the actual compute costs of training a model are relatively unknown unless publicly released, but it been estimated that for GPT-3 for example, the training cost could have been around $1.3M, and that

[14] NVIDIA H100 Tensor Core GPU.
[15] Other examples include Graphcore's IPU (Intelligence Processing Unit), Amazon's Inferentia and Microsoft's Athena
[16] Jouppi, NP, Kurian, G, Li, S, et al (2023), TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings.
[17] Touvron, H, Lavril, Izacard, G, et al (2023), LLaMa: Open and Efficient Foundation Language Models.
[18] Zhao, WX., Zhou, K, et al (2023), A Survey of Large Language Models.
[19] Brown, TB, Mann, B, Ryder, N, Subbiah, M, et al (2020), Language Models are Few-Shot Learners. Number of GPUs and time to train estimated in Narayanan, D, et al (2021), Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM.
[20] Microsoft and NVIDIA. (2012) Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model. Time to train from Deepspeed (2021) DeepSpeed-MoE for NLG: Reducing the training cost of language models by 5 times.

for PaLM the training cost could have been around $10M.[21] Aside from the compute used for pre-training, fine-tuning and inference require further compute.[22] We have heard from stakeholders that compute required for inference can be particularly intensive at scale.

2.19 There are three main ways to access this type and volume of compute: building a data centre, using a publicly available supercomputer or using a cloud computing provider.[23]

Deployment, routes to market and monetisation strategies

2.20 FMs can be developed and released in closed-source or open-source:

(a) ***Open-source models*** are freely shared, and can be used at no cost, subject to their licenses (which can prohibit commercial use).[24] An open-source release can consist of the underlying code, model architecture, and training data, enabling others to replicate the training process. In some cases, it also includes the weights and biases (i.e., the 'knowledge') of the model, such that others can use or fine-tune the model without conducting their own pre-training. Some fine-tuned models have also been made open-source, such that others can use it as trained or conduct additional fine-tuning for their purposes.[25]

(b) ***Closed-source models*** are usually developed privately within companies, and access to the models, as well as information about them, is more controlled and shared only to the extent that the company chooses. These companies may deploy the closed-source models in their own productions or operations, without releasing it externally. Alternatively, they might make them available to external parties to use.

---

[21] Stanford Institute for Human-Centered AI (2023), HAI AI Index Report 2023
[22] Inference refers to each time the model is called upon to make a make a prediction based on new data.
[23] We note that Ofcom is currently conducting a market study of cloud infrastructure services in the UK and intends to publish a final report no later than 5 October 2023. See paragraphs 6.17-6.19 for further details.
[24] We define 'open-source software' as software with source code that anyone can inspect, modify and enhance. We note that the 'open-source model' of software development includes principles such as open collaboration and unrestricted licensing, but we do not refer to those in this report.
[25] We note that, with respect to FMs and FM applications, 'open' and 'open-source' are currently used in a variety of ways. For example, in respect of 'chatbots' like ChatGPT, see Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023, July), Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators for an illustration of several dimensions of openness and one assessment of the varying degrees to which chatbots are open. Similarly, Widder, D. G., West, S., & Whittaker, M. (2023), Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI argue that 'the terms "open" and "open source" are used in confusing and diverse ways, often constituting more aspiration or marketing than technical descriptor and frequently blending concepts from both open source software and open science'. Finally, we are also aware of a process by the Open Source Initiative to define 'Open Source AI'.

2.21 The productisation of FMs is still emerging and evolving. FMs can be deployed by both the owner of the FM and third-parties. Current routes to market for closed-source models include:

(a) ***Integration of FMs into existing products and services*** of the FM owner to improve performance or add new capabilities to existing revenue-generating products and services. In this case, the use of the FM might be considered as a cost of production of these products and services, rather than monetising it directly.

(b) ***Creation of new products or services*** by the FM owner, based on the model. These are usually monetised either through subscription or via a freemium business model (usually consisting of options of free and premium/paid products and services).

(c) ***Providing AI-as-a-service*** enabling third-parties to integrate the model into their own products or services, or to use it for business or personal purposes. There are two main access paradigms:

(i) ***API Access:*** the third party can send prompts to a FM-owner hosted model via an API,[26] and receive a response. These services are often provided on a metered basis (e.g., a price per 1000 tokens of prompt data processed). Usually, in this paradigm the third party has no access to the underlying code, model architecture, training data or model weights and biases. These services can also provide customer-specific fine-tuned models, by using data owned by the third party to fine-tune a personalised model that they can also 'rent' access to via an API.

(ii) ***Model Access***: the third party can deploy the model on their own systems, preventing the need for data sharing with the FM provider. The FM provider may also provide support to the third party to fine-tune the model. These services are also often provided on a metered basis.

2.22 Current routes to market for open-source models include:

(a) ***AI development services*** in which AI labs develop, pre-train and/or fine-tune a model using an open-source model architecture and/or fine-tune an existing pre-trained model and provide the third party with ownership of the model produced (weights & biases).

(b) ***Model hubs or platforms,*** where third-parties can develop and pre-train their own models using open-source model architectures, or fine-tune open-

[26] An Application Programming Interface is code that enables communication between two software programs. In this context, it can be used to submit a prompt to the model and receive the model response in return.

source pre-trained models to develop specialised models for their purposes. Services can include specialised support and computational infrastructure.

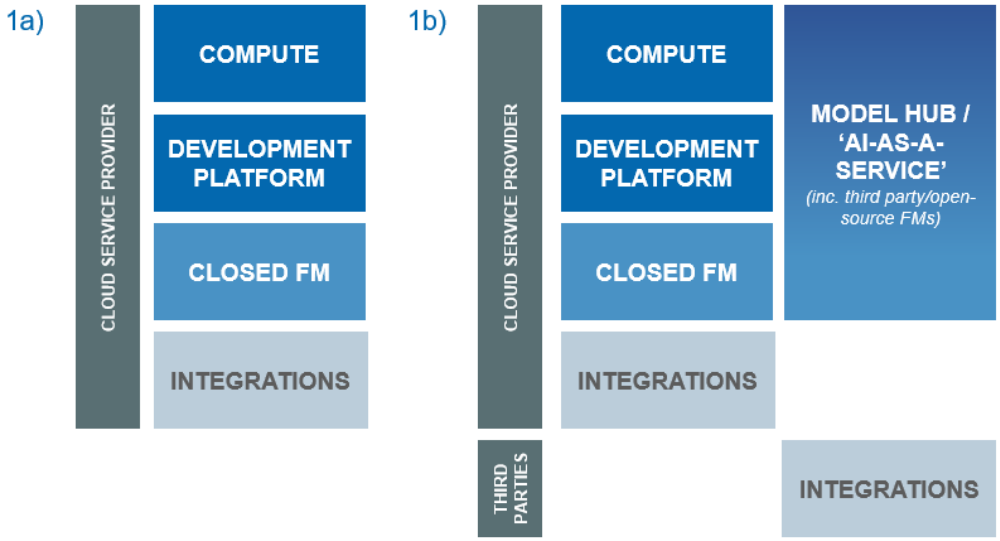
2.23 FM developers may release models open-source, with no direct monetisation, in order to achieve non-commercial objectives. They may also do this to:

(a) Encourage other developers and customers to work with and adopt their FMs (an indirect network effect), which could lead to improvements in the quality of those FMs which they deploy in their other revenue-generating products or operations.

(b) Lower the cost and increase the supply of activities that are complementary to its main product (e.g., firms can sell the compute needed for inference; firms can generate revenue from consultancy and support services for customers using open-sourced models; firms may benefit from or monetise user engagement with more content generated at lower cost due to FMs).

AI supply chains and vertical relationships

2.24 Vertical integration is a type of vertical relationship in which a company controls more than one stage of the production and distribution of a product or service. As described in the above sections, compute infrastructure is a key input for the development of FMs. Several FM developers, such as Microsoft, Amazon and Google, own key infrastructure for producing and distributing FMs such as data centres, servers, network infrastructure and data repositories.[27] Figures 5, 6 and 7 illustrate examples of existing AI supply chains, and the different degrees of vertical integration. In each of these figures, 'Compute' refers to computational infrastructure for FM development, development platforms are additional tools and services provided to developers to aid FM development, and 'Integrations' refer to the integration of FMs into user-facing products and services.



[27] Cobbe, J, Veale, M & Singh, J (2023), Understanding accountability in algorithmic supply chains,

Figure 5: An illustration of two types of vertically integrated supply chains.[28]

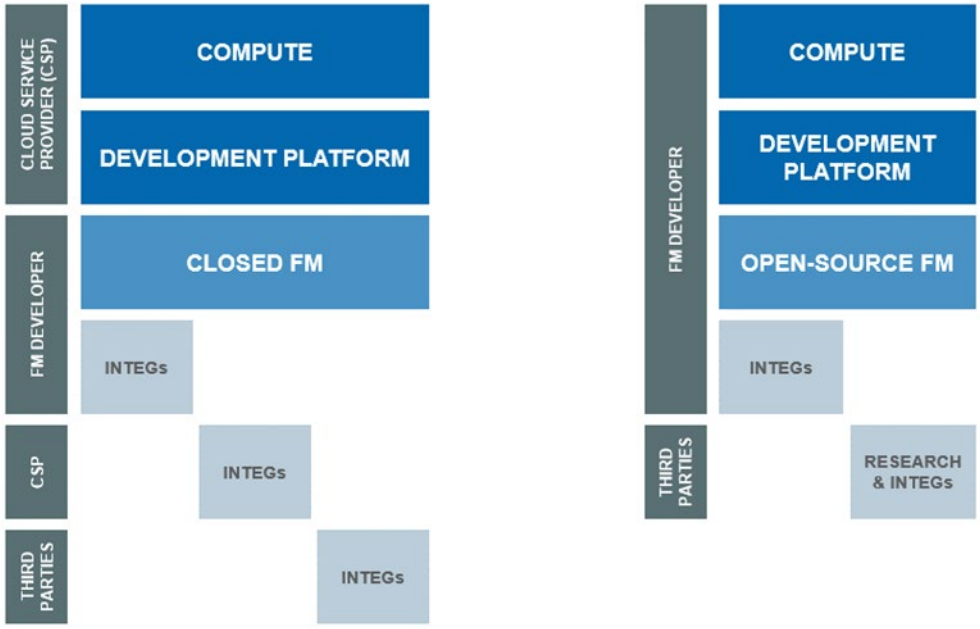


Figure 6: An illustration of two types of partially vertically integrated supply chains.[29]

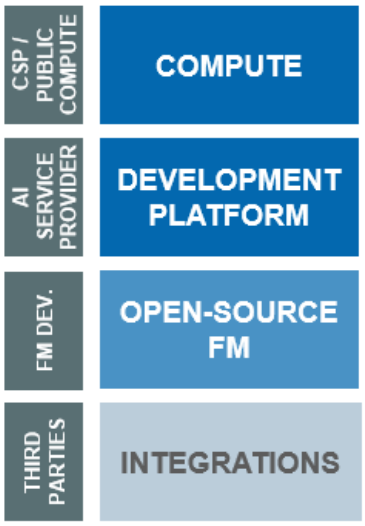


Figure 7: An illustration of a type of fully non-integrated supply chain.

2.25 FM development and deployment is still nascent, but some firms have or may develop ecosystems that are more closed, whilst other supply chains may remain more open.

---

[28] In 1a) the FM developer is a cloud service provider that develops closed-source FMs that are for internal use only. They integrate their FMs into their own products and services. In 1b) the FM developer has an additional supply chain, in which it provides a model hub or 'AI-as-a-service' to customers to integrate FM capabilities into their products and services. These model hubs and services can include provision of third party FMs.

[29] In 2a) the FM developer has a partnership with the cloud service provider to access their computational infrastructure, and the CSP integrates the FM into their products and services. The FM developer also provides the FM to third-parties to integrate into their own products and services. In 2b) the FM developer

2.26 Vertical relationships may also occur when companies use long-term partnerships and strategic investments as an alternative to outright acquisitions and vertical integration. Examples include:

(a) Microsoft's multiyear and multibillion dollar investment in OpenAI, involving a partnership in which Microsoft is the exclusive provider of cloud services for OpenAI's research, products and API services. As part of the agreement, Microsoft is deploying OpenAI's models in consumer and enterprise products, as well as 'new categories of digital experiences'.[30]

(b) Google's partnership with Anthropic, in which Google is providing cloud services to Anthropic to train, scale and deploy its AI systems.[31] Google also participated in multiple funding rounds for Anthropic, including a $300 million Corporate round[32] and participation in a Series C round led by Spark Capital, which raised a total of $450 million.[33]

How FMs are evaluated

2.27 Foundation models are usually evaluated by their developers to assess their performance across a range of tasks, and in some cases to identify harmful or false outputs. Evaluation methods include:

(a) ***Against static datasets*** – whereby various datasets of input-output pairs are used to test a wide range of desirable criteria from accuracy to truthfulness (given the very general ability of these models). A few of the most common collections of these are:

(i) *MMLU*, which measures multitask performance across 57 tasks including maths, history, law and more.[34]

(ii) *BIG-Bench* consists of 204 tasks, contributed by 450 authors across 132 institutions in topics from linguistics to biology and more. It focuses on tasks that are believed to be beyond the current capabilities of LLMs. It also leveraged human raters (more below) to provide a baseline on all tasks.[35]

---

produces open-source FMs that can both be integrated into their products and services, but also may be used at no cost by any third-parties or for research purposes. This developer uses its own infrastructure that is not available to third-parties.

[30] Microsoft Corporate Blogs (2023) Microsoft and OpenAI extend partnership.

[31] Anthropic (2023) Anthropic Partners with Google Cloud.

[32] Crunchbase (2023) Anthropic - Funding, Financials, Valuation & Investors

[33] Anthropic (2023) Anthropic Raises $450 Million in Series C Funding to Scale Reliable AI Products.

[34] Hendrycks, D, Burns, C, et al (2020), Measuring Massive Multitask Language Understanding.

[35] Srivastava, S, Rastogi, A, et al (2022), Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

(iii) *HELM* is a living collection of benchmarks across many tasks, measures multiple metrics at once (eg. accuracy, robustness, calibration, efficiency) for each task, and tries to standardise by using the same evaluation approach (prompting) across all.[36]

(b) ***Model-based*** – whereby one or more other models are leveraged to evaluate the FM. One example would be a language model being evaluated by two other language models working together: a question generation (QG) model and a question answering model (QA) which is given a source reference to use as ground truth. The model under test is then evaluated against the QA model on the same QG questions.

(c) ***Human raters –*** whereby researchers or paid crowd workers are asked or paid to do model-specific evaluation tasks, often creating custom datasets in the process.[37] For example, raters may score the factuality and quality of answers to a set of rater-determined questions. Often this can be quite skilled work, depending on the domain it is being evaluated for. We have been told this approach is usually seen as the gold standard for evaluation. However, since it is model-, task- and methodology-specific, it is hard to then use this to compare performance across models well. Human annotation platforms like Scale and Surge, as well as crowd worker sites like Mechanical Turk and Prolific, are used. We observe that there is screening applied such that only a small proportion of power users are used from the latter.[38]

2.28 One challenge with evaluation is contamination of training data with test/evaluation datasets (e.g. even if an evaluation dataset is filtered out from pretraining, it may be built from a source which is not, such as Wikipedia). This can inflate evaluation scores, given that in this situation, a model may have been trained with the exact information it is being tested on.[39]

2.29 Some firms have started offering external testing and evaluation services too, which can provide additional assurances and enable regulators and other groups that might want independent evaluation services.[40]

---

[36] Liang, P, Bommasani, R, Lee, T, et al (2022), Holistic Evaluation of Language Models.
[37] Datasets built this way are often used to fine tune the model as well, for example in RLHF as discussed in the fine-tuning section.
[38] For example, WebGPT developers filtered contractors by having an undergraduate degree or higher, and further put them through a paid trial, manually checking their work, to only hire the best performers. See Appendix C of the paper: OpenAI (2021), WebGPT: Improving the factual accuracy of language models through web browsing. A similar screening process was implemented for InstructGPT. See Appendix B of Ouyang, L, Wu, J, et al (2022), Training language models to follow instructions with human feedback.
[39] For further information, see for example: BD Tech Talks (2023) Why data contamination is a big issue for LLMs. 17 July. See also Footnote 5 of OpenAI (2023) GPT-4 Technical Report, 27 March, for a real-world example.
[40] For example, Advai uses adversarial testing to stress test models – see Home | Advai Limited.

The FM landscape

2.30 Whilst FMs have become more prominent in public awareness relatively recently, the development of the technology has been evolving since the introduction of the Transformer algorithm by Google in 2017. The first public release of a pre-trained large language model based on the Transformer was the 117M parameter GPT model by OpenAI. Following this, various models of increasing size were developed by many different companies including OpenAI, Google, Meta, Microsoft and Nvidia.

2.31 More recently, in January 2021, image based FMs entered the market with OpenAI's CLIP, an open-source pre-trained model which combines a transformer and neural network to learn mappings between images and text descriptions.[41] OpenAI simultaneously released DALL-E, a proprietary model which uses CLIP to generate images from text prompts. Following this, StabilityAI developed Stable Diffusion based on CLIP and released it open-source.[42] Midjourney also released multiple versions of its proprietary image generator model, launching in July 2022,[43] though details relating to its development remain largely undisclosed. Most recently, in March 2023, Adobe launched its own family of image models called Firefly.[44]

2.32 To date, there has been significant investment into organisations that develop FMs from a range of businesses, including venture capital. The below table illustrates the funding secured by a selection of startups.

Company	Amount of funding raised to date	Prominent investors
OpenAI[45]	$11.3B	Microsoft, Khosla Ventures, A16Z, Sequoia Capital
InflectionAI[46]	$1.5B	NVIDIA, CoreWeave, Microsoft
Anthropic[47]	$1.5B	Google, Spark Capital, Salesforce Ventures, Zoom Ventures
Cohere[48]	$424.9M	Tiger Global Management, Index Ventures, Inovia Capital.
Adept[49]	$415M	Spark Capital, Greylock, General Catalyst, Addition
Stability AI[50]	$89M	Coatue, Lightspeed Venture Partners

[41] OpenAI (2021) CLIP: Connecting text and images
[42] Stability AI (2022) Stable Diffusion Public Release
[43] Midjourney Documentation and User Guide Discord
[44] Adobe (2023) Adobe Unveils Firefly, a Family of new Creative Generative AI
[45] Crunchbase (2023) OpenAI - Funding, Financials, Valuation & Investors
[46] Crunchbase (2023) Inflection AI - Funding, Financials, Valuation & Investors
[47] Crunchbase (2023) Anthropic - Crunchbase Company Profile & Funding
[48] Crunchbase (2023) Cohere - Funding, Financials, Valuation & Investors
[49] Crunchbase (2023) Adept AI - Crunchbase Company Profile & Funding
[50] Crunchbase (2023) Stability AI - Crunchbase Company Profile & Funding

Figure 8: Funding raised to date by a selection of FM startups.

2.33 Whilst the market for FMs is nascent and emerging, there is already a wide range of FMs available through varying levels of access, developed by a variety of different organisations, as illustrated in Figure 9. The range of performance and utility of the FMs listed in Figure 9 is vast, and some models may not be suited to particular use-cases. FMs are currently seeing large gains in performance at each iteration, and many existing models are a 'proof of concept'. Therefore, a significant proportion of FMs developed to date are not currently being used to generate revenue, and may never do so as they become more and more redundant in light of rapid development and evolution. For example, Google previously used its LaMDA family of models to power its Bard chatbot, but that has since been replaced by the more powerful PaLM-2 model.

2.34 Although models that are higher performing over a diverse range of use cases might be considered as 'leading models', Figure 9 does not attempt to distinguish this. We have not found any systematic information, such as share of usage, number of users or revenue, that would enable us to assess which FMs may be market leaders.

2.35 As discussed in paragraph 2.27 there are a range of benchmarks on which the performance of FMs can be assessed and compared. Figure 10 presents a comparison of a selection of models against four benchmarks compiled by the CMA, where scores have been made available by FM developers or other parties. Figure 10 does not contain a complete set of scores for each model, illustrating the current challenges in rigorously comparing model performance. Of the models and scores listed, OpenAI's GPT-4, Google's PaLM-2 and Anthropic's Claude achieve the highest scores.

Organisation / Collaboration	No. of Models	Access Type			Model Names
		Open	*Closed*	*Limited*	
Adept	1	0	1	0	ACT-1
AI2	1	1	0	0	COSMO
AI21 Labs	3	0	0	3	Jurassic-1, Jurassic-1 Instruct, Jurassic-2
Aleph Alpha	1	0	0	1	Luminous
Anthropic	4	1	1	2	Anthropic RLHF models, Claude Instant, Claude, Claude 2
Argonne National Laboratory	1	1	0	0	GenSLM
AssemblyAI	1	0	0	1	Conformer-1
Baidu	2	0	1	1	ERNIE-ViLG, ERNIE-ViLG 2.0
Baidu, PengCheng Laboratory	1	0	1	0	ERNIE 3.0 Titan
Beijing Academy of Artificial Intelligence	1	0	1	0	Wu Dao 2.0
Berkeley	2	2	0	0	OpenLLaMA, Gorilla
BigCode	1	1	0	0	StarCoder
BigScience	4	4	0	0	mT0, T0++, BLOOM, BLOOMZ
Bloomberg	1	0	1	0	BloombergGPT
Cerebras	1	1	0	0	Cerebras-GPT
CMU	1	1	0	0	PolyCoder

Organisation / Collaboration	No. of Models	Access Type			Model Names
		Open	*Closed*	*Limited*	
Cohere	4	0	0	4	Cohere Embed (English), Cohere Base, Cohere Embed (Multilingual), Cohere Command
Databricks	1	1	0	0	Dolly
DeepMind	13	1	12	0	AlphaFold2, RETRO, Gopher, AlphaCode, Flamingo, Gato, GopherCite reward model, GopherCite, Chinchilla, Sparrow Rule reward model, Sparrow Preference reward model, Sparrow, Dramatron
Eleuther AI	1	1	0	0	Pythia
EleutherAI	3	3	0	0	GPT-Neo, GPT-J, GPT-NeoX
Google	38	8	28	2	T5, Internal Google BERT, MUM, LaMDA, GLaM, PaLM, VATT, UL2, Imagen, Parti, Minerva, PaLM-SayCan, MuLan, AudioLM, ViT-e, PaLI, U-PaLM, Flan-U-PaLM, Flan-PaLM, Flan-T5, MultiMedQA, Med-PaLM, w2v-BERT, SoundStream, MusicLM semantic model, MusicLM acoustic model, MusicLM, Phenaki, Noise2Music, Noise2Music pseudolabeler, ViT-22B, Vid2Seq, Flan-UL2, PaLM-E, USM, PaLM 2, Google Joint SLM, Med-PaLM Multimodal
H2O AI	1	1	0	0	h2oGPT
HuggingFace	1	1	0	0	CodeParrot
Inflection AI	1	0	0	1	Inflection-1
Institute of Automation Chinese Academy of Sciences	1	1	0	0	BigTrans
Lehigh University	1	1	0	0	BiomedGPT
Meta	10	7	2	1	FLAVA, OPT, Make-A-Video, ESM-2, Galactica, OPT-IML, LLaMa, SAM, Voicebox, LLaMA 2
Meta, CMU, TTI-Chicago, UC Berkeley, University of Washington	1	1	0	0	InCoder
Microsoft	10	2	6	2	VLMo, Turing NLR-v5, BioGPT, T-ULRv5, Florence, VALL-E, Prometheus, KOSMOS-1, VisualChatGPT, WizardLM
Microsoft, NVIDIA	1	0	0	1	Megatron-Turing NLG
Mosaic	1	1	0	0	MPT
Nanyang Technological University	1	1	0	0	Otter
National University of Singapore	1	1	0	0	GOAT
Naver	1	0	1	0	HyperCLOVA
Neeva	1	0	1	0	Neeva model
NVIDIA	1	0	1	0	Megatron-LM
NVIDIA, Stanford	1	1	0	0	VIMA
OpenAI	18	5	2	11	GPT-2, Jukebox, GPT-3, DALL·E, CLIP, Codex, InstructGPT, DALL·E 2, text-davinci-002, code-davinci-002, VPT, Whisper, text-davinci-003, OpenAI toxicity classifier, Sage, Dragonfly, gpt-3.5-turbo, GPT-4
Salesforce	2	2	0	0	BLIP, CodeGen
Shanghai AI Laboratory	2	2	0	0	InternVideo, Lego-MT
Stability AI	3	3	0	0	Stable Diffusion, StableLM, DeepFloyd IF
Stanford	2	2	0	0	BioMedLM, CORGI
Suno	1	0	0	0	Bark
Together	3	3	0	0	GPT-JT, OpenChatKit moderation model, GPT-NeoXT-Chat-Base
Tsinghua	5	4	0	1	CogView, CogView 2, CogVideo, GLM-130B, CodeGeeX
UAE Technology Innovation Institute	1	1	0	0	Falcon

Organisation / Collaboration	No. of Models	Access Type			Model Names
		Open	*Closed*	*Limited*	
University of Washington	1	1	0	0	Guanaco
Yandex	1	1	0	0	YaLM
You	1	0	1	0	You model
Total	**160**	**68**	**60**	**31**	

Figure 9: An overview of foundation models developed by a range of organisations, illustrating the availability of open, closed and limited access models.[51]

Model Name	*Size*	*Owner*	*Open / Closed* [52]	*MMLU*[53] *(5-shot)*	*BBH*[54] *(3-shot)*	*HellaSwag*[55]	*ARC*[56] *-e / -c*
Llama 2[57]	7B	Meta	Open	45.3	32.6	77.2	75.2 / 45.9
	13B			54.8	39.4	80.7	77.3 / 49.4
	34B			62.6	44.4	83.3	79.4 / 54.5
	70B			68.9	51.2	85.3	80.2 / 57.4
Stable Beluga 2[58]	-	StabilityAI	Open	68.8	-	86.4 (0-shot)	82.7 / 62.0 (0-shot)
Falcon[59]	7B	TII UAE	Open	26.2	28.0	74.1	70.0 / 42.4
	40B			55.4	37.1	83.6	79.2 / 54.5

[51] Open access models are those that can be freely accessed by anyone, whilst closed access models are those kept completely internally to the developing organisation. Limited access models are those for which there is only partial access, such as via an API.. This information in this table may not be exhaustive, was last updated on 16 August 2023 and is derived from research by Stanford University. Stanford University (2023) Ecosystem Graphs for Foundation Models.

[52] There are a variety of ways in which FMs can be more open (including the availability of its code, data, weights, published information and documentation, and the permissiveness of its license), and that the term 'open' and 'open-source' are currently used in a variety of ways to describe FMs. For the purposes of this table, we have emphasised the availability of the model weights to the general public in describing whether each model is 'open' or 'closed'.

[53] Massive Multitask Language Understanding (MMLU) is a benchmark of 57 tasks including elementary mathematics, US history, computer science, law, and more. Hendrycks, D, Burns, C, et al(2020), Measuring Massive Multitask Language Understanding.

[54] Big Bench Hard is a benchmark of 23 challenging language tasks. Suzgun, M, et al (2022), Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them.

[55] HellaSwag is a benchmark based on testing commonsense inference in language models. Zellers, R, et al (2019), HellaSwag: Can a Machine Really Finish Your Sentence?

[56] The AI2 Reasoning Challenge (ARC) is a benchmark comprised of a corpus of natural, grade-school science questions. Clark, P, et al (2018), Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge.

[57] Results from Llama 2 are from Touvron, H, et al (2023), Llama 2: Open Foundation and Fine-Tune Chat Models.

[58] Results for FreeWilly2 are from StabilityAI (2023), Meet Stable Beluga 1 and Stable Beluga 2, Our Large and Mighty Instruction Fine-Tuned Language Models.

[59] Results for Falcon are from HuggingFace (2023) tiiuae/falcon-40b, and Almazrouei et al (2023), Falcon-40B: an open large language model with state-of-the-art performance. The score for BBH was independently assessed by Meta in Touvron et al. (2023), Llama 2: Open Foundation and Fine-Tune Chat Models.

Model Name	*Size*	*Owner*	*Open / Closed* [52]	*MMLU*[53] *(5-shot)*	*BBH*[54] *(3-shot)*	*HellaSwag*[55]	*ARC*[56] *-e / -c*
GPT-4[60]	-	OpenAI	Closed	86.4	-	95.3 (10-shot)	96.3 (25-shot)
GPT-3.5[61]	-	OpenAI	Closed	70.0	-	85.5	85.2
					65.7 (Direct)	86.8 (1-shot)	89.7 / 69.2 (1-shot)
PaLM-2-L[62]	-	Google	Closed	78.3	78.1 (3-shot, CoT)		95.1 (4-shot ARC-C)
Claude 2[63]	-	Anthropic	Closed	78.5 (CoT)	-	-	91.0 (5-shot ARC-C)

Figure 10: Comparison of selected models against benchmarks. Source: CMA.[64]

2.36 Whilst a significant proportion of models in Figure 9 are provided with open access, businesses may be limited in the extent they can use these, as the type of license on which some models are released may restrict their use for commercial purposes. Figure 11 provides an overview of the license types for which the models in Fig 9 were released. The most frequently used 'open' licenses, Apache 2.0, MIT, BSD-3-Clause, and BigScience RAIL v1.0, permit commercial use, although the latter imposes restrictions in order to promote responsible use. Whilst a variety of models are available on these licenses, there may still be limited options for the most performant models. For the majority of models, the license type is unknown, which likely accounts for closed and limited access models.

License	Number of Models
unknown	84
Apache 2.0	35
MIT	13
none	6
BigScience RAIL v1.0	3
BSD-3-Clause	3
Custom / Other	16

[60] Results for GPT-4 and GPT-3.5 are from OpenAI (2023), GPT-4 Technical Report.
[61] Results for GPT-4 and GPT-3.5 are from OpenAI (2023), GPT-4 Technical Report.
[62] Results for the PaLM-2-L model are from Anil, R, et al (2023), PaLM 2 Technical Report.
[63] Results for Claude 2 are from Anthropic (2023), Model Card and Evaluations for Claude Models.
[64] Entries marked '-' indicate this information is unknown.

Figure 11: The number of models with different types of licenses, including those for which the license type is unknown or the model was released without a license (none).[65]

Potential application of FMs

2.37 FMs are already being used across the economy, in a range of early applications, such as:

- **Search**. Microsoft has integrated models from OpenAI into its search engine Bing. Google has announced plans to incorporate generative AI into search.[66] There are also many search and answer engines entering the market such as ChatGPT, You.com and Perplexity.ai.
- **Productivity software.** Google, Microsoft, Adobe, and Slack have all announced plans to integrate generative AI features into their existing products and environments.[67] For example, Adobe has integrated its family of models, Firefly, into Adobe Photoshop, with plans to expand to other software.[68]
- **Finance**. Bloomberg developed its in-house LLM, BloombergGPT, trained on a wide range of financial data.[69] Bloomberg has integrated this into Bloomberg Terminal making it easier for users to write queries and receive financial information.
- **Social media**. Snapchat incorporated the ChatGPT-powered 'My AI' chatbot in its app that replies to users' posts or 'Snaps' with a text based reply.[70]
- **Healthcare**. Generative AI is transforming scientific healthcare and drug discovery, including research on protein folding/expression prediction and rare disease research.[71]

---

[65] Analysis was conducted for the same group of models as Figure 9, derived by the CMA from research from Stanford University. Last updated on 16 August 2023. Stanford University (2023) Ecosystem Graphs for Foundation Models

[66] Google - The Keyword (10/05/2023) How Google is improving Search with Generative AI; Microsoft Bing Blogs (2023): Confirmed: the new Bing runs on OpenAI's GPT-4

[67] Google Workspace (11/05/2023): Introducing Duet AI in Google Workspace; Microsoft (16/03/2023): Introducing Microsoft 365 Copilot ; Abode: AI art generator – Adobe Firefly; Slack (04/052023): Introducing Slack GPT, the future of AI in Slack

[68] Abode News (23/05/2023) Adobe Unveils Future of Creative Cloud With Generative AI as a Creative Co-Pilot in Photoshop

[69] Bloomberg (30/03/2023) Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance

[70] TechCrunch (31/03/2023) Snapchat launches a new generative AI feature, 'My AI Snaps,' for paid subscribers; The Verge (27/02/2023) Snapchat releases 'My AI' chatbot powered by ChatGPT

[71]MIT Technology Review (2022) 10 Breakthrough Technologies 2022: AI for protein folding. 23 February. Avsec, Z, et al (2021), Effective gene expression prediction from sequence by integrating long-range interactions.. Brasil, S, et al (2019), Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter?

- **Robotics.** Researchers have been experimenting with FMs for a range of robotics applications including reasoning, planning, instructions and navigation.[72]

- **Education.** Duolingo, the language learning app, used OpenAI's GPT-4 to create new features such as 'Explain My Answer'.[73]

- **Legal**. Specifio converts patent claims into application drafts. Casetext uses GPT-3 to draft legal briefs.[74]

2.38 Due to the broad range of use cases of FMs, the market for FMs is expected to continue to grow rapidly. For instance, Boston Consulting Group, a management consulting firm, predicts that the total addressable market for uses of generative AI will increase from $18 billion in 2023 to $121 billion in 2027.[75] Furthermore, Gartner, another management consulting firm, predicts that by 2024, 40% of enterprise applications will have embedded conversational AI, up from less than 5% in 2020.[76] By 2025, Gartner expects generative AI to account for 10% of all data produced, up from less than 1% in 2022.[77]

[72] For more information, see GitHub repository GT-RIPL/Awesome-LLM-Robotics for a list of papers experimenting with using FMs for robotics applications.
[73] Duolingo Blog (2023) Introducing Duolingo Max, a learning experience powered by GPT-4.
[74] Sequoia Capital (2022) Expanding on Sequoia's generative AI market map: The 250 companies driving generative AI forward, page 14
[75] BCG Executive Perspectives (2023) BCG Executive Perspectives, The CEO's Roadmap on Generative AI, page 8
[76] Gartner (2023) Generative AI: What Is It, Tools, Models, Applications and Use Cases
[77] Gartner (2023) Top Strategic Technology Trends 2023.

3. Competition and barriers to entry in the development of FMs

Introduction

3.1 This chapter explores competition and barriers to entry in the development of FMs, including the key inputs for the development of FMs and how they are used in FM deployment (this chapter focuses on the inputs outlined in red below, in Figure 12).

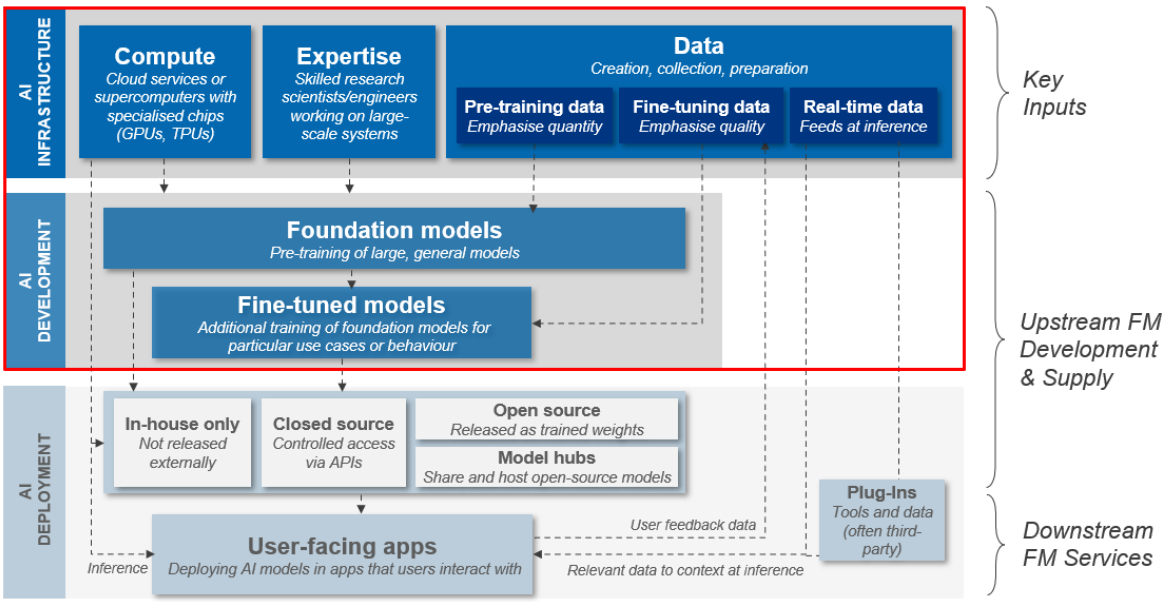


Figure 12: An overview of FM development, training and deployment. This chapter will focus on the parts of the supply chain highlighted by the red box.

3.2 Previous technology-driven markets have shown that network effects and switching barriers can lead to consolidation, weak competition, and a 'winner-takes-most' outcome. This chapter assesses the applicability of these factors to FMs, particularly in the early stage of market development. It also evaluates the likelihood and nature of entry barriers to gauge the potential for market concentration and weak competition.

3.3 It discusses in turn:

- Data requirements;
- Computational resources;
- Technical expertise;
- Access to funding;
- Open-source models; and

- Uncertainties surrounding the future development of FMs, and their potential implications for competition.

Data requirements

3.4 There are different types of data used at each stage of FM development,[78] each of which presents different considerations for examining potential barriers to entry.

Pre-training

3.5 As discussed in chapter 2, FMs are pre-trained using broad and general data sets. We have heard that the volume of pre-training data is a highly important factor in developing a high-performing model, and empirical evidence shows that models are optimally pre-trained by equally scaling both model size and volume of training data.[79] We understand that this has led to a trend of pre-training data sets becoming larger and larger for models developed to date.[80] However, we have also heard that using data of higher quality can provide better performance per token of training data.[81]

3.6 A number of stakeholders said that it is common practice for publicly available data, usually scraped from the web, to comprise the majority of a pre-training data set. For example, a number of high-profile FMs developed to date, including LLaMA (Meta), GPT-3 (OpenAI) and Stable Diffusion (StabilityAI) have been pre-trained entirely on data from freely available sources.[82]

Publicly available data for pre-training could be fully exploited in the next few years, which could increase reliance on proprietary data

3.7 Whilst publicly available data, such as that scraped from the web, can be naturally high quality for pre-training, there appears to be a trend towards adding data from proprietary sources to pre-training data sets. This could be due to concerns that the stock of high quality language data (that could be a source for further performance gains) could be fully exploited within three years.[83] Some stakeholders have explained to us that rather than there necessarily being a hard limit on the availability of data, they are instead seeing significant diminishing returns on the

[78] Outlined on pages 11 and 12 in chapter 2.
[79] Researchers developed a 70B parameter model that achieved higher performance than a larger 280B parameter model by using more pre-training data. Hoffman, J, et al (2022), Training Compute-Optimal Large Language Models.
[80] Hoffman, J, et al (2022), Training Compute-Optimal Large Language Models.
[81] Gao, L, et al (2020), The Pile: An 800GB Dataset of Diverse Text for Language Modelling.
[82] Such as those outlined in para 2.11, page 11 of chapter 2.
[83] Villalobos, P, Sevilla, J, Heim, L, Besiroglu, T, Hobbhahn, M, Ho, A (2022). Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning.

data that is now being added to pre-training data sets from publicly available sources, due to repetition and lower quality.

3.8 Proprietary data may become increasingly important for FM development. We have heard examples of organisations using proprietary sources such as academic journals, image repositories, coding companies and content websites to collate pre-training data.[84] We have also seen reports of data providers increasing prices (Stack Overflow[85], Reddit[86], Twitter[87]), and reports that Microsoft is reducing access to Bing Search index and increasing prices.[88] It is not clear whether certain companies will necessarily have advantages in buying proprietary data from third party data owners, given that there are a range of FM start-ups that have been able to raise significant capital.[89]

Some firms may have data advantages relating to data from activities in other digital markets

3.9 We have heard concerns about the advantages that may derive from owning broad data sets gained from other activities in digital markets.

(a) We heard that access to a web index[90] might be advantageous in the pre-training of FMs, as the web crawling data used to build them can also be used to train FMs. Some stakeholders highlighted that web crawl corpuses, such as C4,[91] have similar utility for pre-training, particularly given that many high performing models have been developed without access to a web index. However, we also heard that search engine providers may have an advantage in obtaining higher quality web crawl data, because:

(i) their crawlers are less likely to be rate-limited[92] or blocked by website owners that want to be discovered and appear on search results pages;

(ii) inherently, web crawl data contains meaningless or useless information (noise), and web indexes provide search engines with the ability to 'find the signal in the noise'. The cost of pre-processing web crawl data to be

---

[84] AP News (2023) ChatGPT-maker OpenAI signs deal with AP to license news stories. Shutterstock (2023) Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data.
[85] Wired (2023) Stack Overflow Will Charge AI Giants for Training Data.
[86] Reddit (18 April 2023) Creating a Healthy Ecosystem for Reddit Data and Reddit Data API Access - Upvoted (redditinc.com).
[87] Twitter Community (2023) Announcing new access tiers for the Twitter API.
[88] The Verge (2023) Microsoft reportedly orders AI chatbot rivals to stop using Bing's search data, 25 March. Tech Monitor (2023) Microsoft hikes the price of Bing API citing AI improvements. 20 February.
[89] See section 2.32 chapter 2 for more detail on funding.
[90] Web indexes are a corpus of web crawl data that has been sorted and organised, enabling search engines to provide rapid results relating to user queries.
[91] See 2.11 chapter 2 for more detail.
[92] The blocking or limiting of activity of users, bots or applications that are deemed to be over-using or abusing a web page.

comparable to a web index was estimated by one stakeholder to be in the order of tens of millions of dollars. Furthermore, we have heard from another stakeholder that filtering web scraped data to find high quality data may be more important than adding data from proprietary sources, given the enormous volumes of data on the web compared to the size of proprietary data sets.

(b) We were told that the data that some companies may own or have easier access to, such as platform interactions and repositories of photos, videos, digital books, audiobooks, music, and podcasts, may have utility in pre-training. In particular, we have seen speculation that YouTube could be an especially valuable source as a repository of 'conversational style' video data with accompanying text data from subtitles and meta-data such as information about the links between videos.[93] However, the utility of these kinds of data may depend on IP and privacy protections. Vertically integrated firms may have easier access to proprietary data, such as data from user interactions on social media. Additionally, proprietary data that could be bought from other sources, such as media and publishing companies or other owners of digital archives, might have similar value for pre-training.

Fine-tuning

3.10 There are two main types of fine-tuning: alignment and domain/task specific.[94] The volume of data required for fine-tuning is typically significantly smaller than for pre-training, with greater emphasis on its quality. A particularly important factor that determines the quality of data for fine-tuning is how well it reflects the specialised 'know-how' of the targeted subject or use case, so that this can be transferred into the knowledge of the model.

Alignment

3.11 The data used for alignment[95] tends to be proprietary and is often human generated (created by humans specifically for the purpose of training) or annotated (labels or additional information are added manually to guide the learning process). This can be achieved in-house using an organisation's employees to generate example conversations, rate model outputs for the purposes of RLHF[96], or by using user feedback data.

---

[93] The Information (2023) Why YouTube Could Give Google an Edge in AI. We note that this source is speculative and that the CMA is open to further views and discussion on this topic.
[94] Outlined in 2.12 chapter 2.
[95] As described in 2.12 (a), chapter 2, alignment is the process of fine-tuning to improve the behaviour of a model to align with the expectations or preferences that a human user may have.
[96] As described in 2.12 (i), chapter 2, reinforcement learning from human feedback (RLHF) trains the model with a reward function that punishes 'bad behaviour'. This relies on human feedback to distinguish between good and bad behaviour, which can be provided by paid contractors or directly from users.

3.12 Alternatively, there is an emerging market of specialist data providers[97] providing high quality labelled data for alignment purposes, including RLHF. There are also efforts to crowd-source data for alignment to share as open-source, including, for example, a human generated and annotated assistant-style conversation corpus named Open Assistant Conversations.[98] However, we heard from one stakeholder that, in general, these efforts may be lagging behind the closed-source provision of data for alignment.

There are differing views on how difficult it is to obtain alignment data

3.13 A number of stakeholders considered alignment data to be costly and difficult to obtain, due to the experience, skill and effort required to annotate data effectively. The complexity and cost of obtaining alignment data may also increase depending on the difficulty of the FM's intended purpose. For example, alignment data for FMs used for medical purposes may require the input of highly skilled professionals. However, others said that crowd-sourced data has fewer issues relating to copyright and privacy than pre-training data, and that the costs of buying alignment data from specialist providers (estimated to be in the range of tens of millions of US dollars) might not be cost prohibitive to VC-backed start-ups.

3.14 Given that user feedback data can also be used to improve a model's behaviour through RLHF, there may be a benefit in having an established user base that can provide large volumes of feedback. However, it is not clear how much of an advantage this confers on providers with more user feedback, given that this data is currently not automatically 'fed' back into the model, but instead requires a rigorous manual review to ensure quality and safety. This means that currently the models are updated sporadically and may provide a practical scaling limit to the amount of user data that can be utilised. On the other hand, one company with a large user base informed us that it had not encountered diminishing returns on the value of user feedback data to date.

There are some emerging open-source alternatives for obtaining alignment data

3.15 An emerging alternative to the use of user feedback data to improve model behaviour is the ability to utilise conversations between humans and existing models that have been shared online. For example, one notable open-source model, Vicuna-13B, was fine-tuned using Meta's pre-trained LLaMA model and a data set of user-shared conversations with ChatGPT collected from a website named ShareGPT. The researchers stated that this model achieved more than 90% of the quality of OpenAI's ChatGPT and Google's Bard, although there are no

[97] For example: Scale AI, Prolific, Surge AI
[98] The Open Assistant Conversations data set is available at: Data Published by Hugging face, Open Assistant/oasst1

universally adopted benchmarks for assessing performance.[99] The long-term viability of this approach may be affected by restrictions within the terms and conditions of leading model services that may limit their use for developing competing machine learning models or technology. For example, such restrictions are currently in place in both Google and OpenAI's terms of use.[100]

3.16 There are also efforts to create and share data sets for alignment fine-tuning which are open to developers.[101] Their effectiveness compared to proprietary data sets is currently unknown, but these data sets are likely to be important for research and development in open-source models.

Domain- or task-specific fine-tuning

3.17 During the domain- or task-specific fine-tuning process, a pre-trained model is trained again on a smaller, specialised data set. Given the vast range of possibilities for domain or task specific fine-tuning, and the variety of different data sources that could be utilised, many businesses may own or have access to data that might have value for fine-tuning a specialised model.

3.18 The likelihood of domain specific fine-tuning data becoming a barrier to entry to develop and deploy a specialised model in any particular sector may also depend on the dynamics of that sector. In particular, the existing distribution of data within certain sectors could influence the ability of industry participants to compete.

Synthetic data

3.19 There are methods to artificially generate more data, called synthetic data, to use for pre-training, fine-tuning, and testing models. Examples include using data from simulations, using existing AI models to generate new data sets,[102] and artificially extending real data. Synthetic data has many benefits as it is less costly to acquire over human-generated data at large scales and it is labelled by design. It also avoids the need for compliance with sensitive or copyrighted data for training models.[103]

3.20 Two studies in which models were trained using FM-generated synthetic data have revealed a risk known as 'model collapse'.[104] These studies show that data generated using existing models can contain defects which pollute future models

---

[99] Vicuna Team (2023) Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality
[100] "You may not use the Services to develop machine learning models or related technology", extracted on 30/06/23 from Generative AI Additional Terms of Service (Google), "You may not... use output from the Services to develop models that compete with OpenAI", extracted on 30/06/23 from Terms of use (OpenAI)
[101] LLMDataHub on GitHub (2023) A quick guide (especially) for trending instruction finetuning datasets
[102] As mentioned in paragraph 3.15, Vicuna was trained on conversations users had with ChatGPT.
[103] IBM Research Blog (2023) What is synthetic data?
[104] Martinez, G, et al (2023), Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet.

in a process that can become irreversible as the learning from original training data is lost. This risk could become more of a concern for FM developers as FM-generated content begins to populate the web and may therefore begin to form a portion of data crawled in future. Given the nascency of this research, it is not yet clear whether this risk will materialise in practice, and what might be done to mitigate it. However, synthetic data might not provide an immediate alternative data source for FM developers if the market moves quickly towards using proprietary data.

Computational resources

3.21 As discussed in paragraphs 2.13 and 2.14 FMs require large, distributed computing systems, often consisting of hundreds of accelerator chips. These are expensive to acquire, have limited availability, and face technical limitations. For instance:

(a) A few businesses produce the necessary inputs, creating dependencies for startups and other companies developing and deploying AI tools. For example, NVIDIA currently leads the market in AI accelerator chips, and Google provides its own TPUs only on GCP.[105] Other businesses which produce GPUs and/or chips for Machine Learning have reportedly struggled to compete.[106]

(b) Initial manufacturing costs are high due to factors including knowledge requirements, expertise, raw materials, highly specialised manufacturing processes, and significant fixed costs. This means that market leaders in semiconductor manufacturing can benefit from economies of scale, which new entrants may find difficult to compete against.

(c) Currently there is a shortage of server GPUs for AI purposes.[107]

(d) Modern GPU architectures encounter bottlenecks with transformers, where they struggle to process certain computations in the model efficiently. This has been exacerbated by the expansion in size of models being developed.[108]

---

[105] CNBC (2023), Meet the $10,000 Nvidia chip powering the race for A.I.,
Reuters (2023) With no big customers named, AMD's AI chip challenge to Nvidia remains uphill fight, and
TPU v4 enables performance, energy and CO2e efficiency gains | Google Cloud Blog.
[106] Reuters (2023) With no big customers named, AMD's AI chip challenge to Nvidia remains uphill fight, and
Bloomberg (31/05/2023), Nvidia Is Soaring. AI Chip Rival Graphcore Can Barely Get Off the Ground.
[107] Hamblen, M (2023), Update: 'Huge' GPU supply shortage due to AI needs, analyst Dylan Patel says, and
New York Times (2023) The A.I. Industry's Desperate Hunt for GPUs Amid a Chip Shortage.
[108] Kim, S, et al (2023), Full Stack Optimization of Transformer Inference: a Survey.

(e) Large technology companies are in different stages of developing their own AI accelerator chips as well. Examples include Amazon AWS Trainium, Meta Training and Inference Accelerator, Microsoft "Athena", and IBM Telum.[109]

Pre-training

3.22 FMs require substantial computing power to pre-train effectively. We have heard that this has, on average, increased in recent years.

3.23 The compute required for pre-training depends on the size and type of model. For example, Meta's LLaMA model (65B parameters) has an estimated compute cost of approximately $4 million, and the relatively large Megatron-Turing NLG (530B parameters) created by Microsoft in partnership with NVIDIA has an estimated compute cost of $100 million.[110] OpenAI reportedly spent over $100 million to develop GPT-4.[111] The computational infrastructure required is one of the reasons downstream firms may choose not to develop their own FM.

3.24 As a model gets larger, the number of operations to train it increases by a power law.[112] This means there is an increasing trade-off between the size of the model and the compute cost to train it. The compute required for inference scales similarly.[113] Additionally, the optimal training data size scales with model size.[114]

3.25 A handful of large technology companies possess substantial compute resources, including AI accelerators, enabling them to pre-train FMs in-house quickly and efficiently.[115]

3.26 We have heard that, with some exceptions, most FM developers would not build the necessary computational infrastructure for pre-training due to the large upfront cost. However, doing so could provide cost savings in the long run for large technology companies or FM developers.

3.27 FM developers that do not already have data centres will generally turn to cloud service providers ('CSPs') for compute. The main options available to FM developers are:

---

[109] AI Accelerator - AWS Trainium - AWS (amazon.com).
MTIA v1: Meta's first-generation AI inference accelerator (facebook.com)
Microsoft Building Codename 'Athena' AI Chip on TSMC's 5nm Node | Extremetech
Servers & Storage (ibm.com)
[110] Towards Data Science (2023) Estimating the Cost of Training LLMs | Towards Data Science, Hugging Face (2021) Large Language Models: A New Moore's Law? (huggingface.co).
[111] WIRED (17/04/2023), OpenAI's CEO Says the Age of Giant AI Models Is Already Over
[112] Hoffman, J, et al (2022), Training Compute-Optimal Large Language Models.
[113] See the glossary at the end of this report for a definition of inference.
[114] Hoffman, J, et al (2022), Training Compute-Optimal Large Language Models.
[115] Ofcom (2023), Cloud services market study: interim report..

(a) Purchase compute resources from CSPs at commercial on-demand rates. This may be cost prohibitive and has no guarantee of availability of the compute necessary for FM training or inference.

(b) Enter an agreement (typically 1 or 3 years) to purchase CSP resources with upfront and/or ongoing costs, at a reduced rate to on-demand prices.[116] This can guarantee that resources will be available, however there can be long lead times (6-9 months).[117]

(c) Enter a commercial partnership with a CSP. This may involve the CSP using the model for its own services[118] or by making the FM available on its cloud services for inference.[119]

3.28 We understand that only a few firms have been able to secure a partnership with a CSP. CSPs may prioritise forming partnerships and allocating scarce computational resources and investment to partners that are more established or where there may be a good strategic rationale to do so. It is likely that those who can secure a partnership will do so (rather than purchase, or enter an agreement to purchase, compute from a CSP), as a partnership provides both a higher-priority access to compute and cheaper rates. Some have even described a compute partnership as necessary or 'critically important' for the development of FMs. The three main CSPs with acceleration capabilities are: Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP).[120] Typically partnerships are achieved through one of these three, although some FM developers maintain relationships with smaller providers or a combination of providers.[121]

3.29 There may be other benefits to commercial partnerships. The CSPs can learn more about the technology behind the FM models and improve their hardware and software accordingly. Access to new accelerator chips can also be beneficial to both the CSP and the developer. The CSP can improve their offering by using the latest hardware as well as mitigate potential supply shortages, while the developer can get access to cutting-edge technology that can help them create more innovative applications. Recent industry developments include Stability AI using

---

[116] AWS, "What are Savings Plans?"
What is Azure savings plans for compute?
Google Cloud, "Committed use discounts for Compute Engine"

[117] Note that different CSPs have different approaches and pricing models to reserving compute power. This is just trying to summarise the trend of compute resources availability.

[118] e.g. Microsoft integrating OpenAI's GPT4 into Bing. Microsoft Bing Blogs (2023), Confirmed: the new Bing runs on OpenAI's GPT-4

[119] AWS deploying Stability AI's Stable Diffusion model on Sagemaker. Stability.ai (2023), Fine-Tune Text-To-Image Stable Diffusion Models With Amazon SageMaker JumpStart

[120] Ofcom (2023), Cloud services market study: interim report.

[121] CoreWeave Partners with EleutherAI & NovelAI to Make Open-Source AI More Accessible — CoreWeave

the AWS Trainium chip to train its next generation of models and Midjourney training its latest models on Google's TPUv4.[122]

3.30 Some stakeholders have expressed concerns regarding the allocation of compute agreements and partnerships, suggesting a potential bias towards entities with pre-existing business relationships or the potential for longer-term commitments. Although select startups can receive investments from CSPs in the form of credits to spend on cloud compute,[123] we have heard a concern that larger companies are more likely to get 'first in line' and make deals to hold larger compute clusters. This can make it hard for smaller organisations such as startups to access the necessary compute to develop and deploy FMs.

3.31 Furthermore, CSP clusters based in the UK have lagged behind other regions in adoption and availability of state-of-the-art accelerators. None of the 3 biggest CSPs have NVIDIA A100s available in the UK, and GCP only offers TPUv4 clusters in the USA.[124] This can be problematic for UK developers working on FMs that need to be trained on sensitive or personal data, as there can be restrictions on storing the data internationally.

3.32 There are some alternatives to renting private compute. For example, Hugging Face used a publicly owned French supercomputer, Jean Zay, in the development of their foundation model BLOOM. Although this has some benefits, such as having full visibility over energy costs, this method cannot be used as flexibly as private compute, as each project requires a research grant. It's unlikely this method is viable for ongoing commercial use, as such Hugging Face also maintains relationships with compute providers.[125] There is also some use of decentralised compute (where idle devices, such as home computers, are used to train models) in academia, but this also not commercially viable currently.[126]

Fine-tuning

3.33 The compute required to fine-tune a FM is orders of magnitude less than for pre-training and is therefore more easily accessible. Fine-tuning a FM can be done on fewer elements of lower grade hardware, for example a single GPU or even a CPU.[127] As fewer and/or less advanced chips are required, the hardware used in

---

[122] AWS Machine Learning Blog (2022) Stability AI builds foundation models on Amazon SageMaker
Google Cloud Blog (2023) Building the most open and innovative AI ecosystem
Google Blog (2023) Google's Cloud TPU v4 provides exaFLOPS-scale ML with industry-leading efficiency
[123] AWS Activate for Startups, Founders, & Entrepreneurs (amazon.com)
AI startup program | Google Cloud
[124] The Economist (2023), How to make Britain's AI dreams reality; AWS, Amazon EC2 P4 Instances; Google Cloud - Cloud TPU pricing; Google Cloud - Location Accelerators
[125] Hugging Face Blog (2023), Hugging Face and AWS partner to make AI more accessible.
[126] Yuan, B, et al (2023), Decentralized Training of Foundation Models in Heterogenous Environments.
[127] IEEE Spectrum (2023) The Case for Running AI on CPUs Isn't Dead Yet - IEEE Spectrum.
Church, KW, Chen, Z & Ma, Y (2021), Emerging trends: A gentle introduction to fine-tuning, Natural Language Engineering, Volume 27, Issue 6, pp. 763 – 778.

fine-tuning is cheaper and more widely available. We have not heard concerns of issues in access to compute referring only to the fine-tuning stage of development.

3.34 The compute requirements for fine-tuning may also continue to decrease. Recent technological innovations in fine-tuning which build on top of open-source models (such as EleutherAI's GPT-J and Meta's LLaMA model) through novel techniques such as Low-Rank Adaption (LoRA),[128] allow the fine-tuning of a model at a fraction of the cost and time. Some examples of these include Vicuna-13B[129], Alpaca[130] and Koala[131] with fine-tuning costs of $300, $600 and $100 respectively. However, this process relies on access to a pre-trained model to iterate on top of. The ability for new models developed in this way to remain competitive could depend on the continued release of high performing and freely accessible pre-trained models to fine-tune.

Inference

3.35 Compute is also required each time the model does inference.[132] A single inference requires very little compute, so this is accessible and sustainable at small scales. However, as the model size and/or number of users increases, the compute required to run the model for inference also increases, and inference at scale can still require large amounts of compute. This can require large clusters of accelerators to manage demand and reduce time taken to infer (latency).[133] Hence, unlike some other digital markets, there is a non-negligible marginal cost associated with deployed FMs.

3.36 FM developers can deploy models themselves. This requires them to develop and maintain the infrastructure to facilitate the inference, as well as pay for the cost of computing it. This can be through an API, where users can programmatically query the model and receive the inference as a response. Examples of FM APIs include those provided by OpenAI and Anthropic.[134]

3.37 Alternatively, some cloud service providers can provide FM inference (and fine-tuning) through their platforms and APIs. For example, the Amazon Bedrock API

---

[128] Developed by Microsoft Hu, EJ, et al (2021), LoRA: Low-Rank Adaptation of Large Language Models.
[129] Vicuna-13B is a an open-source chatbot trained by fine-tuning the LLaMA pre-trained model on user-shared conversations collected from ShareGPT.
[130] Stanford HAI (2023) Alpaca: A Strong, Replicable Instruction-Following Model
[131] Koala is a Dialogue Model trained using publicly available data created through supervised fine-tuning from the LLaMA pre-trained model. When compared against ChatGPT measuring real human preferences, more than 50% of the time users either prefer Koala or have no preference.
[132] As set out in the glossary, an inference is each time the model is called upon to make a make a prediction based on new data.
[133] Coreweave, Serving Inference for LLMs: A Case Study with NVIDIA Triton Inference Server and Eleuther AI
[134] OpenAI Blog (2020), We're releasing an API for accessing new AI models developed by OpenAI.; Product \ Anthropic

allows access to models from AI21, Anthropic, and Stability AI.[135] Azure has a similar service with OpenAI models.[136] Amazon also offers FMs through SageMaker, a platform that allows more granular development than an API.[137] It is often the CSP who manages the deployment cost of FMs in this instance.

3.38 We have not heard specific concerns about being able to access compute for inference through these or other methods although it is possible that this market feature could benefit vertically integrated firms or firms with vertical relationships who likely face lower inference costs.

Technical expertise

3.39 FMs are complex and require a high level of technical expertise to develop and train. The technical expertise required includes cutting-edge knowledge of machine learning, as well as significant practical expertise in data engineering and high-performance computing. Many job advertisements for FM developers include data scientists, machine learning (ML) engineers, NLP/computer vision experts. These will also require highly skilled individuals, usually requiring Masters/PhD qualifications.

3.40 A decade ago, most advanced ML models came from academic research. The development of such models has since shifted to industry, which is likely due to access to greater resources and talent.[138] A significant amount of prominent research has been carried out and published by industry, e.g. Google published the first paper on transformer models; Microsoft published the LoRA paper.[139] However, there have been models developed with public funding, in academia[140] and by non-profit firms[141] at both pre-training and fine-tuning stages.

3.41 There has been a shift of talent from academia to industry with data scientists and ML/AI researchers. For example, 65% of new AI PhDs were hired by industry in 2021, compared to 41% in 2011.[142] This may be because jobs in the sector fetch a

---

[135] Foundation Model API Service – Amazon Bedrock – AWS
[136] Azure OpenAI Service – Advanced Language Models | Microsoft Azure
[137] AWS, JumpStart Foundation Models
[138] Stanford Institute for Human-Centered AI (2023), HAI AI Index Report 2023.
[139] Vaswani, A, et al (2017), Attention is all you need.
Hu, EJ, et al (2021), LoRA: Low-Rank Adaptation of Large Language Models.
[140] The Falcon model was developed by the publicly funded Technology Innovation Institute in UAE Falcon LLM - Home (tii.ae)
BLOOM was developed with many academic collaborators, with funding for the publicly owned Jean Zay supercomputer Founding members (notion.site)
The Alpaca model was developed by Stanford CRFM Stanford CRFM.
[141] RedPyjama models were developed by Together Together.ai Blog, Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models
StableLM models were developed by Stability AI Stability AI Launches the First of its StableLM Suite of Language Models — Stability AI
[142] Stanford Institute for Human-Centered AI (2023), HAI AI Index Report 2023.

high premium for developing and researching new models, with which we were told that academia cannot compete.[143]

3.42 FMs require generic skills in engineering to deploy at scale, as well as specialised skills to pre-train. The latter is scarce as only existing engineers in the sector have had access to the necessary infrastructure which is limited to industry. Some firms are also training experienced software engineers to work with AI. This means that larger firms may be able to acquire this talent more easily, either by hiring those with the relevant skills with high salaries, or by investing to train existing software engineers in the relevant skills.

3.43 No stakeholders have raised concerns about non-compete clauses or publication restrictions imposed on employees during this review.

Access to funding

3.44 The cost of training and deploying FMs is significant, especially for pre-training. Only a limited number of organisations have the resources to do so independently without securing additional funding.

3.45 The evidence we have seen shows that currently smaller players are able to secure funding from investors.[144] This has led to a notable increase in FMs in recent years, with an estimated 160 FM since 2018.[145] As outlined in the background section of this report (paragraph 2.32), companies such as Adept, Cohere, and Stability AI have successfully obtained funding to develop and deploy FMs. Google and Microsoft have also provided funding to various FM developers, including Anthropic[146] and OpenAI.[147] A French start-up, Mistral AI, was able to secure $113 million in seed funding after only four weeks of existence.[148]

3.46 The long-term impact of this funding on the growth of smaller players in the sector remains uncertain. Nevertheless, it is evident that such funding plays a crucial role in enabling these players to establish a presence in the sector.

---

[143] AI Index Report 2023 – Artificial Intelligence Index (stanford.edu)
[144] See Background section at paragraph 2.31.
[145] See Background Figure 9.
[146] It is reported that Anthropic has received a total of $450 million in funding from Google. See here: Reuters (2023) Google-backed Anthropic raises $450 mln in latest AI funding
[147] Microsoft has invested a total of $13 billion in OpenAI over three rounds of funding. The first round, in July 2019, was for $1 billion. The second round, in January 2021, was for $1.5 billion. And the third round, in January 2023, was for $10 billion. See here: Reuters (2023) Microsoft to invest more in OpenAI as tech race heats up
[148] TechCrunch (2023), France's Mistral AI blows in with a $113M seed round at a $260M valuation to take on OpenAI

Open-source models

Pre-training

3.47 As shown in Figure 9, many firms have kept their highest-performing pre-trained FMs closed-source, keeping the model weights (the internal 'knowledge' of the model) a trade secret, and providing access via an API or through user-facing applications.

3.48 A number of pre-trained models have been released as open-source as illustrated in Figure 9 chapter 2.

3.49 Open-source models are typically more transparent and accessible than closed-source models. This is because the code and parameters of open-source models are publicly available, which makes it easier for researchers and developers to understand and improve them. Closed-source models, on the other hand, are less transparent because companies may choose not to release the full details of their models or how they were trained.

3.50 The greater transparency of open-source models has several benefits. Users can have a better understanding of how the models work, which can help them to assess their accuracy and reliability. They can also modify the code of open-source models to improve them or add new features. Additionally, users can contribute to the development of open-source models by submitting bug fixes or new features.

3.51 The development of open-source models can also be crowdsourced. Users can make suggestions to the original developers to improve the model. They can also fork the model[149] and make their own improvements, increasing the availability and diversity of FMs, without affecting the original model.

3.52 There are also risks associated with open-source models. They have a decentralised nature, where contributors come from diverse backgrounds and have varying levels of expertise. This can make it challenging to establish and enforce consistent governance policies across the entire open-source community. There is also the challenge of monitoring their use by bad actors who intend to use them for harmful reasons.[150]

3.53 According to some stakeholders and reporting, open-source pre-trained models are generally smaller and perform less well than the highest-performing closed-source models. This may be because closed-source models are typically trained

[149] When a model is forked, it typically means that someone takes an existing model, and creates a separate version of it. The forked model becomes an independent entity that can be modified, fine-tuned, or further developed without affecting the original model.
[150] See Chapter 5 for a fuller discussion of consumer protection issues relating to the use of FMs.

on larger data sets, with more powerful hardware, and optimised for enhanced performance.

3.54 However, the extent and implications of any gap in capabilities between open-source and closed-source models are still uncertain, and it remains to be seen to what extent any such gap will be maintained over time.[151]

Fine-tuning

3.55 Developments in the fine-tuning of open-source models have led to significant breakthroughs in their capabilities. For example, some fine-tuned open-source LLMs have claimed to reach a comparable performance to fine-tuned closed-source models, such as ChatGPT.[152] For example, In March 2023, the 13-billion parameter Vicuna LLM model claimed to deliver 90% of ChatGPT's quality.[153]

3.56 The LLM leader boards published by Hugging Face also show the progression made by open-source models, with the highest performing models achieving increasingly impressive performance metrics.[154] However, the credibility of some of the performance metrics used to rate performance of both open- and closed-source models has been questioned,[155] given that they are highly subjective, and developers are inclined to choose metrics that are most favourable to their models.

3.57 With access to open-source pre-trained models, researchers and developers have used a number of fine-tuning methods such as parameter efficient fine-tuning ('PEFT') and low rank adaption ('LoRA'). These reduce the resources required to fine-tune a pre-trained model, enabling quick iteration in a cost-effective manner.[156]

Uncertainties

3.58 In addition to FM developers needing access to computing power, data, technical expertise, and capital to compete effectively, our analysis has also identified a number of key uncertainties regarding the future development of FMs. These uncertainties could lead to positive or more concerning outcomes for competition and consumers:

- Will access to proprietary data become necessary to compete?

---

[151] This is explored more in the Uncertainties section below.
[152] LMSYS ORG (2023) Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.
[153] See here: EcoAGI (19/08/2023), Vicuna-13B: An Open-Source ChatGPT Alternative That Impresses GPT-4
[154] Data Published by Hugging Face, Open LLM Leaderboard
[155] Semianalysis (2023) Google "We Have No Moat, And Neither Does OpenAI" (semianalysis.com).
[156] Semianalysis (2023) Google "We Have No Moat, And Neither Does OpenAI" (semianalysis.com).

- Will models become larger?
- Will FMs be highly generalised?
- Is cutting edge performance required to compete?
- Will large technology companies and first movers have an advantage?
- Will open-source models remain a key part of the market?

3.59 Whilst the development of FMs is still in its early stages, there are a number of different ways that these models could evolve in the future. For example, FM development could become centred on a small number of large companies or on a large number of small and medium-sized companies.

3.60 A **positive market outcome** for consumers, businesses and the wider economy would arise if there were multiple independent firms competing with one another to produce leading FM models, with innovative firms able to access the inputs they need to enter, expand and compete effectively. In that scenario, firms would be able to experiment with different business models and forms of monetisation, including the supply of FMs on both an open-source and closed-source basis so others can continue to build on existing FM capabilities.

3.61 However, a **concerning market outcome** could emerge if access to inputs is restricted so only a handful of firms can create and maintain the leading models. As a result, those remaining firms would develop positions of strength which could give them the ability and incentive to only provide models on a closed-source basis and make them subject to unfair prices and terms.

3.62 In the following section, we outline the key uncertainties[157] and analyse their potential impact on competition and market outcomes, both positive and concerning. There may also be other uncertainties not discussed here that could impact competition and consumers.

Will access to proprietary data become necessary to compete?

3.63 As highlighted in paragraph 3.7, the stock of publicly available, high-quality data for developing FMs may soon have been fully exploited for gaining improved model performance. If this is true, it is likely that improved performance will need

[157] Given the inherent unpredictability of the future it would likely be impossible to create a comprehensive and accurate set of possible outcomes. We have not attempted to do so. Instead, these are necessarily stylised outcomes, and we do not claim that any of these options will materialise in the way we describe or at all. Rather they merely extrapolate, for analytic purposes, market features and trends that we think may emerge based on the evidence we have seen that we consider could have an impact on competition and consumers.

to be achieved through different means, but it is currently uncertain how this might play out.

3.64 Identifying new training methods and model architectures to find efficiencies or improved performance is currently an important area of research for FM developers. However, it is not clear to what extent, and how quickly, these innovations may occur. Access to large volumes of high-quality training data may confer less of an advantage if new methods to achieve increased performance, using fewer resources, emerge and become accessible to a range of competitors in the near term. However, access could become more of an advantage if innovations in efficiency do not keep pace with performance improvements achieved from continuing to increase the volume of training data.

3.65 There are also uncertainties about what the most effective way to gain access to increasingly large volumes of training data will be. As discussed in paragraph 3.9(a), access to a web index might improve the ability of a FM developer to obtain or identify high quality data from web crawls. Therefore, access to a web index could become necessary to compete effectively in this situation.

3.66 However, if obtaining large volumes of data from proprietary sources (in addition to data available from web crawling) becomes necessary to develop the most competitive models, then access to proprietary data could become a key factor that influences competition. On the one hand, this could stimulate a dynamic market of data providers who supply data on fair and equal terms to a range of FM developers. But, on the other hand, if the most useful sources of proprietary data for training are only accessible to a small range of existing FM developers, for example due to their activities in other digital markets, this could potentially stifle competition.

3.67 A factor that could influence this uncertainty is the potential enforcement of copyright law relating to the use of web crawled training data. For example, this has been the subject of various ongoing legal action, such as those filed by Getty against StabilityAI[158] and authors Mona Awad and Paul Tremblay against OpenAI.[159] Should courts rule in favour of the plaintiffs in cases such as these, this could reduce the amount of data available for training, or increase its price, which in turn could amplify the potential advantages of owning proprietary data (see paragraphs 6.48 and 6.49).

[158] Marks & Clerk (2023) Getty Images taking UK action against Stability AI for copyright infringement in AI training. Thomson Reuters (2023) Getty Images V Stability AI. Reuters (2023) Lawsuits accuse AI content creators of misusing copyrighted work.
[159] The Guardian (2023) Authors file a lawsuit against OpenAI for unlawfully 'ingesting' their books

Will models become larger?

3.68 FMs have trended to becoming larger. One of the first transformer models released was BERT in 2018 which had 354 million parameters.[160] Since then, PaLM, GPT-3, and Megatron-Turing NLG have been developed with hundreds of billions of parameters.[161] Even high ranking open-source LLMs have at least tens of billions.[162]

3.69 The principal reason behind this trend is an observed positive relationship between scale (size of model, amount of training data, amount of compute and training time) and performance, known as 'scaling laws'.[163]

3.70 However, there are increasing trade-offs between model size and the cost of computational requirements for training and inference.

3.71 If FMs continue to enlarge to stay competitive, then the development of these models may be restricted to firms that already have access to the required computational infrastructure. They would also need increasingly larger data sets for training, which has its own uncertainties, as discussed above. This arms race scenario could result in a concentrated market, with the inputs of compute and data creating high barriers to entry and/or expansion for many entrants.

3.72 A concentration in the development of pre-trained models may not necessarily exclude a diverse market for the fine-tuning of FMs. Lower requirements for compute to conduct fine-tuning may encourage developers to take advantage of "off-the-shelf" FMs and fine-tune them to their own, or a client's needs.

3.73 There is uncertainty about the extent to which performance can be gained from increasing model size, partly due to a lack of standardised performance metrics on which to assess this. There are also risks of "inverse scaling", where models may lose performance as they get larger than a threshold.[164]

3.74 Alternatively, models may become smaller. Some incentives to develop smaller FMs include:

---

[160] Devlin, J, Chang, MW, Kenton, L & Toutanova, K (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[161] Google Research (2022), Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

Brown, TB, Mann, B, Ryder, N, Subbiah, M, et al (2020), Language Models are Few-Shot Learners.

Shoeybi, M, et al (2020), Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism.

[162] Data Published by Hugging Face, Open LLM Leaderboard

[163] Kaplan, J, et al (2020), Scaling Laws for Neural Language Models.

[164] It was found in a recent study that as models increase in scale, they may become more inaccurate. One explanation given is that larger models tend to have stronger priors, memorised phrases which are learned from the training data and are chosen over the prompt context. These priors can sometimes lead the model to make incorrect predictions, especially when the given context is different from the training data. For more information, see: McKenzie, IR, et al (2023), Inverse Scaling: When Bigger Isn't Better.

(a) Smaller models require less upfront cost to pre-train, as they have less computation and data required.

(b) They are more sustainable in the long-term, as their inference costs go down with model size.

(c) Performance may begin to plateau as models grow, hence developers could build more efficient models to increase profitability.

3.75 There are a range of methods by which models can become smaller and more efficient. For example, sparsification[165] and quantization[166] reduce the computational requirements. Other potential innovations in model architecture and training techniques could lead to more efficient scaling laws while maintaining similar performance.

3.76 Models can also be compressed after their pre-training. While this does not reduce the computational cost of training them, it does make them cheaper to deploy. This can have advantages for the provider as their inference cost is reduced, possibly to zero if the inference is computed at the end user's device. Methods include pruning (a form of sparsifying after training) and quantization.[167]

3.77 Alternatively, knowledge distillation is a method where a large, trained model (teacher) is used to train a smaller model (student) to mimic it as closely as possible.[168] For example, it has been reported that Google used a distilled model of PaLM to run on a smartphone.[169]

3.78 If FMs continue to become increasingly large to improve their performance, this may raise barriers to entry and reduce competition between FM developers. This could happen if users require the highest performance models (see 'Is cutting edge performance required to compete?') and the primary means to reach that is through increasing the model size. By combining the increase in development

---

[165] Sparsification is a method that uses sparse matrices in the model, which contain many zero values. This can reduce the storage and computational resources to train a model without loss of accuracy.
Dao, T, et al, (2021), Pixelated Butterfly: Simple and Efficient Sparse training for Neural Network Models/ Zhou, A, et al (2023), Learning N:M Fine-grained Structured Sparse Neural Networks From Scratch.
[166] Quantization uses smaller, less precise, number formats such as 8-bit floating points for model parameters. This can retain performance while reduce the computational cost for training. However, current accelerator chips are designed to use 16-bit (and higher) floating points so this would require developments in hardware.
Micikevicius, P, et al (2022), FP8 Formats for Deep Learning.
NVIDIA A100 (NVIDIA)
Heyman, K (2023), 'Will Floating Point 8 Solve AI/ML Overhead?', Semiconductor Engineering
[167] A Fast Post-Training Pruning Framework for Transformers (neurips.cc)
Bondarenko, Y, Nagel, M, & Blankevoort, T (2021), Understanding the Challenges of Efficient Transformer Quantization. https://arxiv.org/abs/2302.14017Kim, S, et al (2023), Full Stack Optimization of Transformer Inference: a Survey.
[168] Towards Data Science (2021) Distillation of BERT-Like Models: The Theory | by Remi Ouazan Reboul | Towards Data Science.
[169] Financial Times (2023) The race to bring generative AI to mobile devices | Financial Times (ft.com).

cost, and the potential for computational resources to continue to be limited, this would likely result in higher barriers to entry. However, barriers might be reduced through possible innovations in development of FMs, such as reducing the size of models and increasing the training efficiency without sacrificing performance.

Will FMs be highly generalised?

3.79 As a general-purpose technology, FMs can be deployed in a number of products and services. For example, FMs can be used to power chatbots, generate text, or translate languages. We have heard that models might develop to a point where they no longer require significant customisation to become highly effective in a wide range of tasks. This could reduce the need for domain specific fine-tuning, and lead to an increase in the use of other techniques, like prompt engineering[170] and retrieval augmentation.[171]

3.80 If FM development were to advance to the point where the most powerful models are highly effective for a wide range of tasks without requiring extensive customisation, then this could have an impact on market dynamics. For example, the number of FMs available could consolidate, as a small number of models could meet the needs of most users. This consolidation would be a concern if it reduced the incentive for FM developers to compete and innovate. This is because a decrease in competition could reduce the demand for new and different models. This could lead to stagnation in FM innovation, as developers would have less incentive to develop FMs or improve existing models.

3.81 However, other stakeholders believe that FMs will continue to be most effective when fine-tuned for specific tasks. In this situation we might see a proliferation of specialised models that are less intensive to produce being developed or fine-tuned by a wide range of organisations making using of their domain-specific data. In this situation we might see a proliferation of specialised models that are less intensive to produce being developed or fine-tuned by a wide range of organisations making using of their domain-specific data.

3.82 It is also possible that a mixture of general FMs and task-specific fine-tuned models will emerge, with different downstream markets subject to different approaches.

---

[170] Prompt engineering is the process of creating effective prompts that enable AI models to generate responses based on given inputs. Prompts are essentially instructions that tell the model what to do. See Liu, P, et al (2021), Pre-train, Prompt and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.

[171] In retrieval augmentation, a language model is first given a prompt. The model then retrieves relevant information from a knowledge base and uses this information to generate its response. This helps to improve the accuracy and relevancy of the model's response. See Guu, K, et al (2020), REALM: Retrieval-Augmented Language Model Pre-Training.

Is cutting edge performance required to compete?

3.83 Currently the FMs at the cutting edge in terms of performance are those using vast amounts of inputs. However, it is possible that open-source or closed-source models will not need to achieve a comparable performance level to the highest performing models to act as a competitive constraint.

3.84 One of the reasons for this is because not all potential applications for FMs may require cutting-edge performance. For instance, certain tasks such as classifying customer reviews or generating text descriptions for products, among others, could be effectively accomplished with smaller models or those fine-tuned for the specific purpose. One example of this type of model is Einstein GPT by Salesforce which is a model which focuses on Customer Relationship Management (CRM). For the task of CRM, this model could give better performance than a model at the 'cutting edge', due to its task specific nature.[172]

3.85 Currently the highest performing models are closed-source. As discussed in paragraph 3.53 the performance of pre-trained open-source models currently lags behind proprietary ones. However, with some fine-tuning, open-source models could be competitive with some closed-source models. If models do not have to compete at the cutting edge to be utilised in some applications, this could lead to competition from both smaller or fine-tuned closed-source models or open-source alternatives.

3.86 If models at different levels of performance are utilised in some applications, there could be lower barriers to entry in model development, as the production of such models may require less compute, less expertise and potentially different data than at the cutting edge (for example some models may require less data to become proficient to the required level, or a model for a specific task may require domain specific data which could potentially be easily accessed). This could result in a more competitive market outcome than if all applications look to use models which are at the cutting edge.

3.87 This could be an area where open-source models have the potential to provide a competitive constraint to closed-source models, even if in the future open-source is unable to achieve cutting edge performance.

3.88 If cutting edge models become required or preferred across the majority of applications, it may be the case that this performance can only be achieved by one or a small number of models at the frontier. In this situation we could end up with a concentrated FM market dominated by a few large players. However, if cutting edge performance does not require large models with lots of inputs due to possible

[172] News & Insights (2023), Why Einstein GPT Marks the Next Big Milestone in Salesforce's AI Journey

future innovations, some of which are described in 3.74, this could make the cutting edge accessible to a broader spectrum of developers.

3.89 The development of more generalised models may influence this dynamic. If models develop such that they can compete in multiple specialised markets (this may not have to be at the cutting-edge), specialised models may have to be closer to the cutting-edge to act as a competitive constraint.

Will large technology companies and first movers have an advantage over others?

3.90 Large technology companies' access to vast amounts of data and resources may provide them with an insurmountable advantage over smaller organisations, making it hard for them to compete. However, the extent of this advantage is uncertain, as it depends on a number of factors, including economies of scale, economies of scope, and feedback effects. The FM landscape is also evolving, and the long-term impact of smaller, closed or open-source models is yet to be fully understood. Factors such as innovation, community collaboration, and emerging technologies could potentially disrupt the existing dynamics, making the future competitive landscape less predictable.

3.91 In this section, we examine the potential impact of first mover advantages, economies of scale, economies of scope and feedback effects in more detail. If these effects are very strong, it could lead some firms to develop strong positions that are difficult to contest in future.

3.92 As discussed in paragraph 2.30, the first public release of a FM was by OpenAI, followed by various models developed by Google, Meta, Microsoft and NVIDIA.

3.93 These early movers in the FM market have the potential to enjoy several benefits. First, investing early in the development of FMs may offer lower input costs before demand drives these costs up (for example the cost of compute and data). Second, they can establish themselves as leading providers of these models, gaining an advantage in terms of brand recognition and customer loyalty.[173] They may also have more time to experiment and refine their models, giving them an edge in terms of performance and capabilities, or potentially be able to exploit a larger amount of user data with which to re-train their models. Lastly, early entry provides an opportunity to build a robust ecosystem of partners and developers, resulting in a broad user network and a stronger market position.

3.94 Being early to release major FMs does not guarantee success or the ability to capitalise on this advantage.[174] Mistakes made by early entrants can be learning

[174] Fast followers may be able to capitalise on the developments made by first movers by leveraging their early investments in research and development. They can enter the market later and avoid the high costs of innovation, which can often lead to market success.

opportunities for others, such as violating intellectual property laws or failing to comply with privacy regulations. Early movers may also develop less accurate or efficient FMs. Stranded investments are another risk, where resources that have already been invested in become obsolete as the market evolves.[175] Additionally, pre-trained open-source FMs may allow for new entrants to catch up quickly. This means that later entrants can benefit from early movers' work without having to invest the time and resources that they did. This can make it difficult for first movers to maintain a competitive advantage.

3.95 Other factors, such as competitors' ability to catch up in terms of model performance and capabilities, will also play a crucial role in determining whether first movers will be able to establish and sustain a competitive edge.

3.96 Economies of scale, economies of scope, and feedback effects (learning and network effects) are also potential advantages for some larger firms. Large technology companies that develop FMs may have an advantage due to economies of scale[176], as the initial high training costs can be offset by efficiently using trained models to train multiple others at a significantly reduced cost.

3.97 Economies of scope[177] may also offer advantages, with large technology companies able to leverage hardware, algorithms, and training techniques across multiple models. The leveraging of these shared resources may allow large technology companies to maximise their resources and enhance overall performance in a more streamlined and cost-effective manner.

3.98 Learning effects from user-generated data, such as feedback conveyed through features like thumbs up and thumbs down buttons in chat interfaces, may also provide an advantage for large technology companies. While this data may not directly improve the FM from which it is collected, it can be used during pre-training or fine-tuning of future iterations, giving an edge to providers with access to such data. If this data proves valuable, those with large quantities of user feedback data may be able to use it to improve their models significantly.

3.99 It is not clear that currently deployed FMs benefit from network effects (although, as described below, there may be network effects in the context of some downstream applications such as plug-ins). For example, one stakeholder argued that currently once a FM completes its pre-training or fine-tuning, its performance level is essentially fixed,[178] with the number of users having no immediate direct

---

[175] For example, where a company allocates significant resources towards the training of a FM that becomes outdated or obsolete before its deployment or commercialisation.

[176] Economies of scale refer to the cost advantages that a company can achieve by increasing the scale of production. As the level of production increases, the average cost of producing each unit decreases.

[177] Economies of scope refer to the cost advantages that a company can achieve by producing a variety of products or services together rather than producing them separately.

[178] This might change if development and training cycles get shorter, or if FM providers are able to implement safely some form of active, continual 'online' learning.

impact on user experience. Nevertheless, those with large user bases may benefit from data feedback loops in developing future models as discussed in paragraph 3.14. The extent to which this could give early movers an advantage depends on the value of this feedback data.

3.100 A market participant has also questioned the significance of economies of scale, economies of scope and feedback effects, arguing that some providers of open-source models, which may operate without the benefit of these advantages, can compete effectively with providers of closed-source models.[179]

3.101 We have also heard some speculative views about other mechanisms that could give rise to positive feedback effects. We have heard that FMs can be used as coding assistants, making them useful even to the teams building further FMs. It is uncertain the extent to which, and in which domains and use-cases, FMs will be able to benefit from recursive self-improvement (the idea that FMs could be used to improve their own ability to improve), including by using 'self-play'[180] approaches and using trained models to generate data for subsequent iterations of training (see the section on 'synthetic data' above).

Will open-source models remain a key part of the market?

3.102 There has been significant and rapid innovation occurring in open-source FMs, focusing on the efficient development of high performing, fine-tuned models. Many developers are taking advantage of pre-trained models that have been made publicly available, thus eliminating the need for the large upfront costs of an FM. This provides options in the market that are not dependent on commercial FMs. However, it is not clear whether fine-tuned open-source models will remain competitive with closed models in the longer-term.

3.103 There is some uncertainty around the ability for open-source developers to secure funding to develop pre-trained FMs. This could depend on the appetite of investors to support the open-source movement, and also on the ability of open-source developers to commercialise some part of their business models, such as selling software along with an open-source model which makes interacting with or fine-tuning the model easier. Creating some sort of return through this channel could make open-source a more attractive investment.

---

[179] In addition, a leaked memo from a Google engineer suggests that open-source models pose a competitive constraint on closed alternatives. This is because open-source models, in the view of the memo's author, has various advantages including being more cost-effective, customisable, and have fewer usage restrictions. See Semianalysis (2023) Google "We Have No Moat, And Neither Does OpenAI" (semianalysis.com)

[180] Self-play is a technique for improving the performance of reinforcement learning agents, and was used by DeepMind's AlphaZero to play games like chess and go. (See Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140-1144.)

3.104 Another potential concern is whether high quality, competitive open-source pre-trained models will continue to be released. For example, we have heard from a prominent contributor to open-source pre-trained models who has indicated its intention to discontinue the development of FMs after releasing several smaller-scale open-source language models.

3.105 Some more 'open' FMs (i.e., in a narrow sense that model weights are publicly available) may nevertheless be subject to licensing terms that can limit their use and distribution. For example, Meta's original LLaMA model, from which a lot of advances in open-source development have been made, could only be used for research purposes,[181] excluding commercial applications. OpenAI's terms of use[182] may also prohibit the use of its models to develop competing models. These licensing terms raise important questions about who will control access to 'open' and open-source models in the future and what limits they may put in place on how they are developed and used.

3.106 Therefore, as closed-source pre-trained models advance in terms of cutting-edge capabilities, the ability of models fine-tuned from open-source alternatives to remain competitive becomes uncertain. However, there have been developments in this area. Meta[183] has released LLaMA 2, which can be licensed for commercial use, albeit with some restrictions.[184] OpenAI[185] has also announced plans to release open-source FMs that can be used for commercial purposes in the near future. These developments could have a significant impact on the development and use of FMs, as these models may be accessible to a wide range of developers and businesses.

3.107 Finally, we have also heard an additional potential concern that suppliers of FMs could initially use open-source models to develop their ecosystem of partners and developers, but later choose to transition away from open-source approach.[186]

3.108 Another challenge is striking the right balance between innovation and maintaining ethical standards. Open-source development encourages experimentation and creativity, which is a driving force behind its success. However, this also means

---

[181] Kan, M (2023), Meta Debuts AI Language Model, But It's Only for Researchers
[182] See 2(c)(iii) here: Terms of use (openai.com)
[183] Meta and Microsoft Introduce the Next Generation of Llama | Meta (fb.com)
[184] LLaMA 2 has been release with certain restrictions, such as excluding licensees with over 700 million active monthly users and limiting its outputs' use for improving other LLMs. See the terms and conditions on Meta AI Request access to the next version of Llama (accessed 6 September 2023).
[185] Zhang, M (2023), OpenAI Readies Open-Source Model as Competition Intensifies.
[186] See The Federal Trade Commission's blog of 29 June 2023 (Generative AI Raises Competition Concerns | Federal Trade Commission (ftc.gov)) which states 'Experience has also shown how firms can use "open first, closed later" tactics in ways that undermine long-term competition. Firms that initially use open-source to draw business, establish steady streams of data, and accrue scale advantages can later close off their ecosystem to lock-in customers and lock-out competition.' See also Widder, D. G., West, S., & Whittaker, M. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI, which makes similar arguments.

that there is the possibility of misuse or unethical practices, even when safeguards are put in place.[187] This may make enforcing standards complex as, unlike closed models where the control rests with a single entity, open-source projects rely on voluntary contributions from a wide range of individuals and organisations.

Conclusion

3.109 FM developers need access to the key inputs of computing power, data, technical expertise, and capital to compete effectively. These inputs are important for developing, deploying, and using FMs. If access to these key inputs were to be constrained, then FM developers may not be able to compete with larger, more established businesses that have greater resources. This could lead to a decrease in competition and innovation in the FM sector, which could ultimately harm consumers.

3.110 We have also identified a number of key uncertainties regarding the future development of FMs. The impact of these uncertainties on competition in the FM sector is not yet known, but they could have an adverse impact if they manifest in the following ways:

- If proprietary data becomes important for training FMs, a lack of access to this data could create a barrier to entry and expansion for smaller organisations or research groups. This could prevent these entities from effectively competing with larger and more established players.
- If models need to become larger to keep up with the increasingly complex needs of applications, it could further disadvantage smaller organisations with limited computational resources and infrastructure.
- Similarly, if cutting-edge performance is necessary to be competitive, it could place a significant burden on smaller entities that may not have the resources or expertise to achieve such advancements.
- If FMs become highly effective in a wide range of tasks, the number of FMs could consolidate. This is because a small number of highly effective models could meet the needs of most users. This consolidation could reduce the incentive for FM developers to compete and innovate, as there would be less demand for new and different models.
- Large technology companies' access to vast amounts of data and resources may allow them to leverage economies of scale, economies of scope, and

[187] Europol (2023), ChatGPT - the impact of Large Language Models on Law Enforcement

feedback effects to gain an insurmountable advantage over smaller organisations, making it hard for them to compete.

- There are potential challenges to the continued development and deployment of open-source models. These include possible licensing restrictions, funding uncertainty, and the potential for closed models to outperform open-source models in the longer-term.

3.111 If the FM sector develops in this way, it could make it harder for some firms to compete effectively. This could stifle innovation, limit diversity in approaches, and impede the dynamic nature of the market that has been instrumental in the advancement of FMs thus far.

Given the likely importance of FMs across the economy, we would be concerned if access to the key inputs required to develop FMs were unduly restricted, in particular restrictions on data or computing power. The market is more likely to trend towards positive outcomes if:

- A range of FM developers can access the key inputs they need to build FMs, including data, computing power, capital and expertise, on fair commercial terms without undue restrictions.
- Initial successful FM developers face an ongoing competitive constraint from new entrants, so they do not gain an entrenched and disproportionate advantage by being an early mover in the market, having economies of scale or benefiting from feedback loops.
- There are a range of models – including open-source and closed-source models and FMs pushing at the frontier of new capabilities – available for firms to choose from.
- Firms are unable to use their leading position in other markets to unduly restrict access to firms they compete with in those markets or other competing FM developers.

4. The impact of FMs on competition in other markets

Introduction

4.1 In the previous chapter, we looked at the key inputs to FMs and upstream FM development and supply. This chapter focuses on the deployment of FMs in consumer facing applications, products, and services ("downstream FM services"), at the next level in the value chain, as shown in the red box in Figure 13 below.

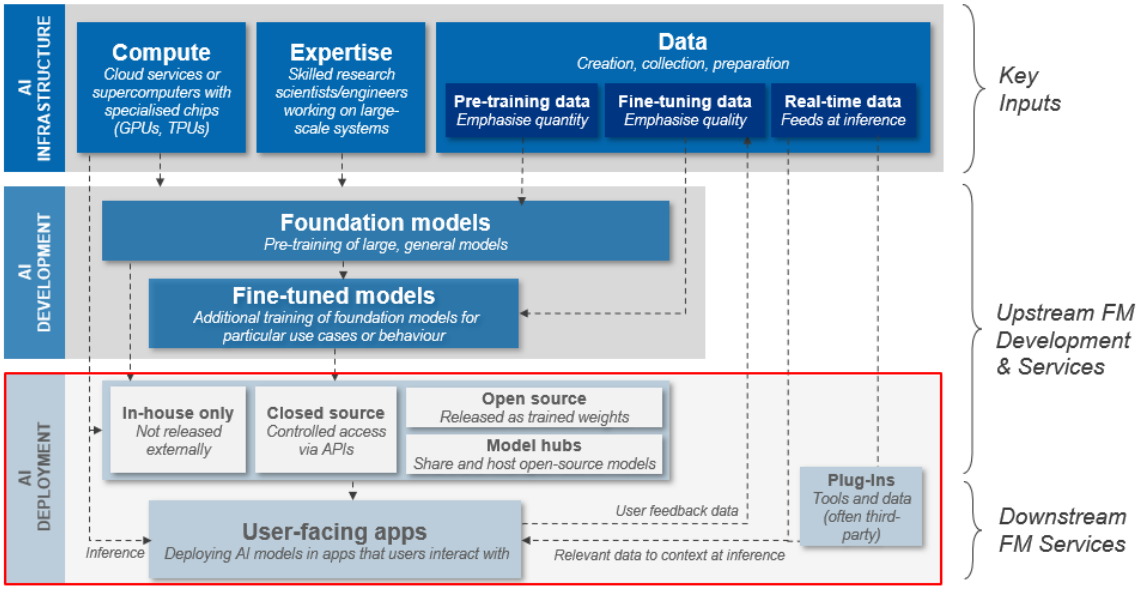


Figure 13: Deploying FMs in downstream FM services.

4.2 The development of FMs and their deployment (i.e. the upstream and downstream levels) are closely linked, and developments in either part of the value chain can impact upon competition on the other. We will therefore refer to the upstream FM development and supply where it is relevant to assessing competitive outcomes at the downstream level.

4.3 In this chapter, we will consider:

- How firms can deploy FMs in downstream markets;
- The potential impact of FMs on competition in downstream markets, including potential risks from vertical integration; and
- The key uncertainties about how FM products and services may develop that could determine the impact FMs have on competition in downstream markets.

Deploying FMs in downstream markets

FMs could become an important input in a wide range of markets

4.4 A wide range of third parties told us that they expect AI to be applied in many different downstream contexts and our review has identified several areas where FMs already are, or are likely to become, an important input in future.

4.5 As discussed in paragraphs 2.21 and 2.37 of the Background section, consumers can interact with FMs in a variety of ways. For instance, some are deployed as standalone services such as chatbots like Anthropic's Claude.[188] Others are integrated within existing services such as Bing.[189] Others are add-ons within existing applications and services such as Duolingo.[190]

4.6 Incorporation of FMs into downstream services could allow firms of many shapes and sizes to be more productive and efficient. In some markets, this can make it easier for a wider range of businesses to develop products and services. Examples include creative industries, such as marketing, where FMs are being used by firms – large and small - to produce materials, such as visuals, which may allow them to compete more effectively with market incumbents.

4.7 On the other hand, they could also create new or entrench existing positions of market power for the firms that develop that product or service. This may be more likely where the firm that develops the new product or service has market power in related markets, potentially enabling firms to leverage that market power to help acquire market power in a new market and/or to lock customers into broader ecosystems of related products and services from that firm.

4.8 As illustrative examples of downstream deployment, we discuss in the case studies below how firms are using FMs in search and productivity software services.

Firms can access FMs in a number of ways

4.9 FM providers can make FMs available using a variety of deployment options. Similarly, downstream firms can access or source FMs in several ways, and they can choose one or a combination of these options. For example, a firm can:

- Develop a FM in-house from scratch – where the firm takes responsibility for creating and maintaining the FM and applying it to their own products and services. This option can provide the firm with full control of the FM and the

[188] Anthropic \ Claude 2
[189] The Official Microsoft Blog (2023) Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web
[190] Duolingo blog (2023), Introducing Duolingo Max, a learning experience powered by GPT-4

flexibility to adapt it for its own needs. However, it can be expensive and time-consuming.[191] Firms that adopt this approach include Bloomberg, Pfizer, and Adobe.[192]

- Partner with a FM provider to enhance an existing FM – the firm may use a third party FM, which removes the need to develop the model itself, but fine-tune the model with its own proprietary data to tailor it to its business needs.[193] This offers some flexibility and potential for differentiation because the downstream firm owns and fully controls the fine-tuned FM.[194] Whilst this is cheaper than developing an in-house FM, it may still be expensive, time-consuming and require technical expertise.[195]

- Buy API access to a third party FM and FM deployment tools[196] - the firm could buy API access to a FM and FM deployment tools. This is often much cheaper and faster than in-house development.[197] There are several third party FMs and deployment tools to choose from and it is relatively easy to switch between them. Third parties told us that downstream firms can also try different models before committing to their use. On the other hand, this option offers limited flexibility and makes downstream firms reliant on a third party supplier.[198] Firms that adopt this approach include Duolingo, Shutterstock and Expedia.[199]

- Provide a third party plug-in – plug-ins are another way for firms to add FM capabilities to their services. As shown in Figure 14 below, a firm can develop a plug-in (the plug-in provider) to augment its offering with an FM-

---

[191] A BCG March 2023 report estimated that it costs $**50-90m**+ to create a new, cutting-edge FM for complex models (BCG Executive Perspectives, The CEO's Roadmap on Generative AI, March 2023, page 13). Note that the costs may change rapidly.

[192] Bloomberg (30/03/2023), Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance; Pfizer Doubles Down on AI/ML to Bring Transformative Medicines to Patients | BioSpace ; Adobe Blog (2023) Bringing Generative AI into Creative Cloud with Adobe Firefly

[193] See paragraphs 3.17- 3.18 above.

[194] Owning and controlling the fine-tuned FM might mean the developer has access to the model weights. One advantage is that the developer can adjust the model weights to the particular use case. See paragraphs 3.17- 3.18 above.

[195] A BCG March 2023 report estimates that it costs **$1m-10m**+ to support fine-tuning complex proprietary data (BCG Executive Perspectives, The CEO's Roadmap on Generative AI, March 2023, page 13). Further, we heard that the same fine-tuning technical costs are also needed to keep in line with newer versions of the pre-trained model.

[196] See paragraph 4.9 above for discussion of pre-trained FMs and FM deployment tools.

[197] Based on a BCG March 2023 estimate, costs range from **$10k-$100k** to fine-tune existing FM for related tasks (e.g.if you were to fine-tune ChatGPT for writing medical articles). (BCG Executive Perspectives, The CEO's Roadmap on Generative AI, March 2023, page 13.)

[198] There is limited flexibility because the developer does not get access to the fine-tuned model, only controlled API access to the model outputs.

[199] Shutterstock (2022), SHUTTERSTOCK PARTNERS WITH OPENAI AND LEADS THE WAY TO BRING AI-GENERATED CONTENT TO ALL; Expedia Group (2023), CHATGPT WROTE THIS PRESS RELEASE — NO, IT DIDN'T, BUT IT CAN NOW ASSIST WITH TRAVEL PLANNING IN THE EXPEDIA APP; Duolingo Team (2023), Introducing Duolingo Max, a learning experience powered by GPT-4.

based service (the plug-in host), such as ChatGPT. This allows the third party to add an advanced natural language interface to its service.

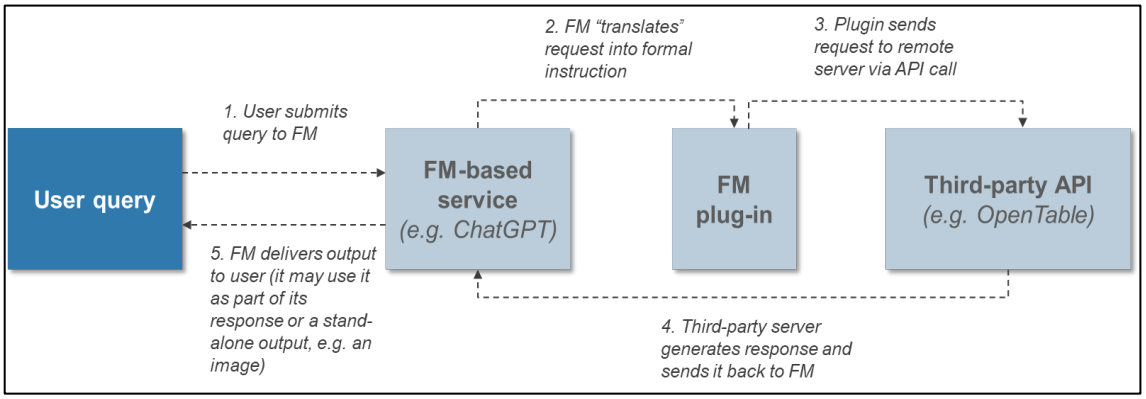


Figure 14: How a FM plug-in responds to a consumer query

- We understand that firms share some information with the plug-in host, as indicated by the information flows in Figure 14. For example, a consumer asks, using natural language, whether there are any good Italian restaurants in the area. The FM-based service will translate this query to a specific instruction which will be sent, via the plug-in, to a third party server. The server will then return a response (e.g., a list of restaurants), which the FM-based service will use to respond to the consumer. The extent of data shared between FM-based service and the third party, as well as how it is used will depend on the terms agreed between them in each case.

- Plug-ins offer a quick, easy and cheaper way to benefit from FM functionality without needing to fine-tune or build in-house. Two stakeholders told us that plug-in providers can currently easily multi-home and provide multiple plug-ins. Some firms that provide plug-ins include Klarna, Expedia and OpenTable.[200]

4.10 In summary, downstream firms can currently choose from a wide range of deployment options between which they can easily switch or multi-home. We think this is one of the most important drivers of FM deployment and competition in downstream FM services, as further discussed below.

Firms are monetising FM services in different ways

4.11 Firms are monetising downstream FM services in different ways. Some firms are currently offering FM-based innovations at no extra charge to the consumer. For example, OpenAI currently offers its chatbot answer engine, ChatGPT, for free.[201] Other firms require consumers to pay an additional charge for FM-integration in

[200] ChatGPT plugins (openai.com)
[201] Introducing ChatGPT (openai.com)

existing products, e.g. Duolingo charges customers a fee to access its FM-based 'Explain My Answer' feature.[202]

4.12 Some firms are also offering both a free and a paid version of their FM service. For example, OpenAI offers ChatGPT to consumers for free, but is also piloting a subscription version, ChatGPT Plus, for $20/month. The subscription version offers added benefits such as peak time access, faster response times and priority access to new features.[203]

4.13 Some firms may also seek to monetise FM services indirectly e.g. in an upstream or adjacent market. Andreessen Horowitz, a VC firm, notes that firms may mainly be monetising FM services through inputs such as cloud and estimate about 10-20% of revenue from generative AI services goes to cloud companies.[204]

4.14 It is too early to tell which monetisation approach firms will settle on and they may pursue a variety of different approaches. The emerging picture on monetisation will have important implications for competition in downstream markets, as further discussed below.[205]

Potential impact of FMs on competition in downstream markets, including potential risks from vertical integration

4.15 We now consider how:

- FMs could drive competition and disrupt incumbent firms;
- Vertical integration and partnerships could affect competition in downstream FM services; and
- Features of downstream markets could affect competition in upstream FM development.

How FMs could drive competition and disrupt incumbent firms

4.16 At the downstream level, FMs have the potential to transform the ways people and businesses use software, creating entire categories of new products and adding distinctive capabilities to existing ones. Today, many start-ups are trying to compete effectively with incumbent firms, using FMs to differentiate their products.[206] For example, Whisp is an FM powered fact-checking service that can instantly verify claims, news articles, and other online information. Journalists and

---

[202] Malik, A, (2023), Duolingo launches new subscription tier with access to AI tutor powered by GPT-4
[203] Introducing ChatGPT Plus (openai.com)
[204] Bornstein, M, Appenzeller, G, Casado, M (2023), Who Owns the Generative AI Platform?
[205] See the discussion in paragraphs 4.50 and 4.53 below.
[206] See e.g. Generative AI startups list | Dealroom.co and Khan, J (2023), Some small startups making headway on generative A.I.'s biggest challenges

news organisations can use it to carry out real-time analysis of facts used in reporting, and social networks can use Whisp to detect harmful and misleading content.[207] Another example of a potential disruptor is Notion, a text editor trying to differentiate itself through its integrated AI writing assistant.[208]

4.17 Even if new entrants are ultimately unsuccessful, the threat of entry and disruption in contestable markets could provide competitive discipline to incumbent firms. One third party told us that a broad set of competitors are emerging, and they are trying to stay ahead of them, stressing the importance of differentiation as a key factor to remain competitive.

4.18 New FM services have the potential to disrupt even long-standing market positions. For example, OpenAI's ChatGPT chatbot has the potential to disrupt Google's position in search. But it is also possible that things go the other way, and FMs reinforce existing incumbent firms' market positions. For instance, the adoption of FMs by leading search engine providers could strengthen their positions in online search, as they may be best placed to develop and implement these new technologies effectively.

4.19 Deploying FMs could drive competition in new and existing markets as firms learn more about FMs and deploy them in downstream services. However, it is unclear how the competitive dynamics in FM services will evolve and the extent to which they will disrupt current market positions.

4.20 In the case studies below, we look at search and productivity software as illustrative examples of downstream FM deployment respectively. They cover how firms are using FMs in search and productivity software services and their potential impact on competition.

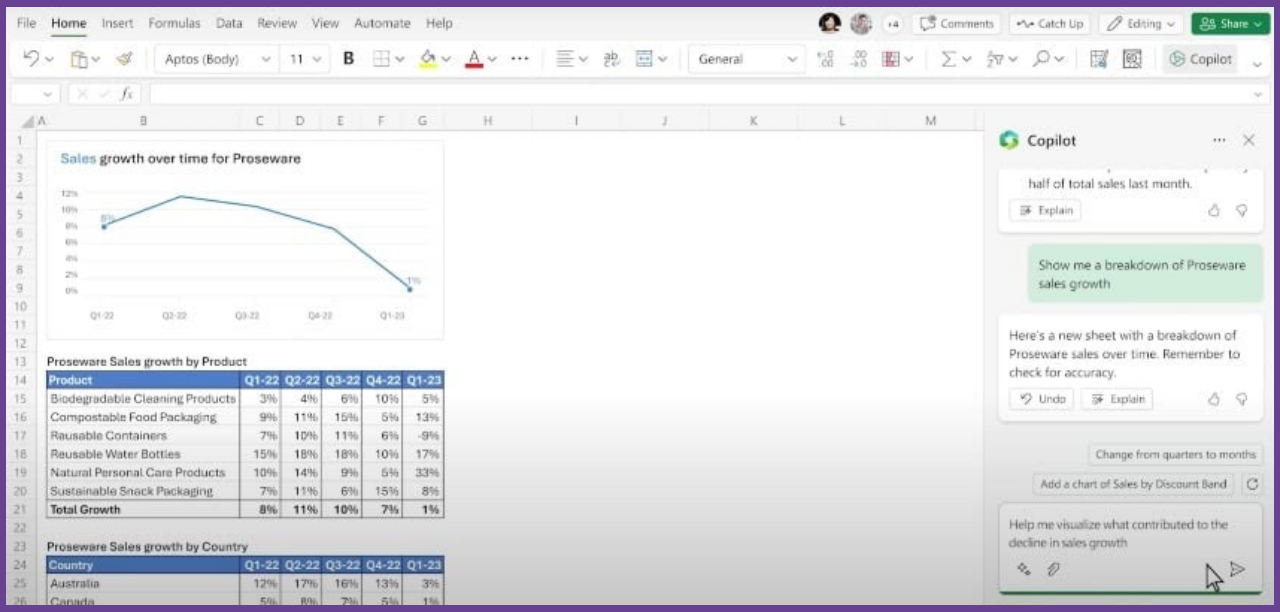
Case study: Productivity software

From information we have received and observations we have made of several firms beginning to integrate generative AI into their products, FMs appear to be an increasingly important input to productivity software services. A third party said the benefits it was seeing include "increasing productivity, unleashing creativity and lessening the drudgery of work".

[207] Whisp — Instant, AI-powered fact-checking
[208] Notion AI | Work faster. Write better. Think bigger.

Google, Microsoft, Adobe, and Slack all announced plans to integrate generative AI features into their existing productivity software services.[209]

For example, Microsoft is testing a Copilot virtual assistant that allows users to seamlessly access Outlook email, Office suite, Bing search, Azure cloud services as well as a range of third-party plug-in providers. The Copilot can draw on materials across these services, e.g. create a presentation combining data from a Bing search and the user's own Office-created documents.



Source: Microsoft[210]

Figure 15: shows how Microsoft's Copilot can respond to natural language inputs to analyse data in Excel.

Further, Google recently announced an experimental product, NotebookLM, that is an FM powered 'virtual research assistant' that can summarise facts, explain complex ideas and help brainstorm, all personalised to the consumer's own notes and sources.[211]

Examples of start-ups offering FM powered productivity software services include the following:

- Notion AI offers AI-powered services for organising notes, and highlighting action items and key takeaways.[212]
- Jasper is an AI platform for businesses that helps creators use generative AI to "break through writer's block". It can be trained on a particular brand to generate content in tailored formats, tones and languages.[213]

---

209 Microsoft 365 Blog (2023) Introducing Microsoft 365 Copilot—A whole new way to work; Google Workspace Blog (2023), A new era for AI and Google Workspace; Generative Fill - Adobe Photoshop ; Introducing Slack GPT, the future of AI in Slack | Slack

210 Microsoft 365 Copilot in Excel - YouTube

211 NotebookLM: How to try Google's experimental AI-first notebook (blog.google)

212 Notion AI | Work faster. Write better. Think bigger.

213 Plans & Pricing - Jasper

- Ask AI, a FM chatbot by Codeway Dijital, can write stories, poems and scripts, as well as help with language practice and text translation.[214]

The potential impact of FMs on competition in productivity software is unclear and will depend on several factors. For example:

- Google, Microsoft, Adobe, and Slack are all integrating FM-based features into their existing productivity software services. This may give them a competitive advantage over new entrants because they can direct their existing consumer base to new FM-based features and use data from their existing services to develop those features.
- FM powered productivity software services could evolve towards customised ecosystems integrated with multiple other adjacent FM services e.g. search. These could offer increased convenience and utility to consumers. However, it may also make it harder for firms offering standalone FM services to compete, especially if it is difficult for consumers to switch away from these ecosystems.
- The incentives and behaviours of firms providing FM powered productivity software with their own FM model will also depend on factors such as how FM powered features will be monetised.

We discuss each of these uncertainties in the sections below.

Case study: Search

Firms are deploying FMs in search in different ways.[215] Some are using FMs to add new features and improve functionalities in existing search engines, whilst others are deploying FMs to create new services such as chatbot answer engines.

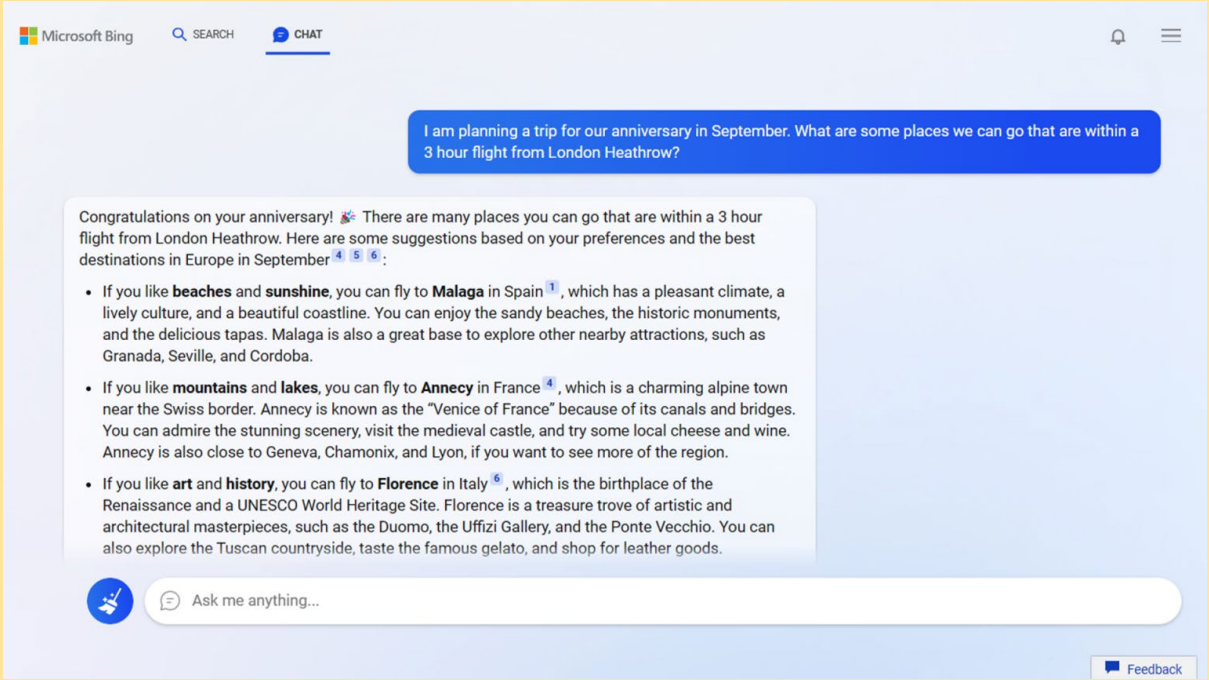
For example, OpenAI launched its chatbot answer engine, ChatGPT, in November 2022. ChatGPT instantly appealed to consumers with its ability to respond to questions in a conversational, human-like way and compose written content including articles, social media posts, essays and code.[216]

Firms have also deployed FMs in existing search services. For example, Microsoft launched a new, FM powered version of its Bing search engine in February 2023.[217] The new Bing deploys GPT-4 and seeks to create a 'unified experience' by integrating (1) Bing search engine; (2) Bing Chat, Microsoft's chatbot answer engine; and, (3) Microsoft Edge browser.[218] Consumers can access Bing Chat via its general search engine by clicking on the 'chat' button, as shown in Figure 16 below. Microsoft's Edge browser also features a new Bing sidebar, as seen in Figure 17.

---

[214] Ask AI - Chat with Chatbot - Apps on Google Play
[215] In this report the term 'search' includes both FM-powered search engines (like new Bing) and chatbot answer engines (like ChatGPT). It is not intended to capture a formal market definition.
[216] Introducing ChatGPT (openai.com); Hetler, A (2023), What is Generative AI? Everything you need to know.
[217] Your AI-Powered Copilot for the Web (microsoft.com); Lardinois, F (2023), Microsoft launches the new Bing, with ChatGPT built in
[218] Bing Chat | Microsoft Edge. Bing chat is also integrated into Microsoft Edge's sidebar.

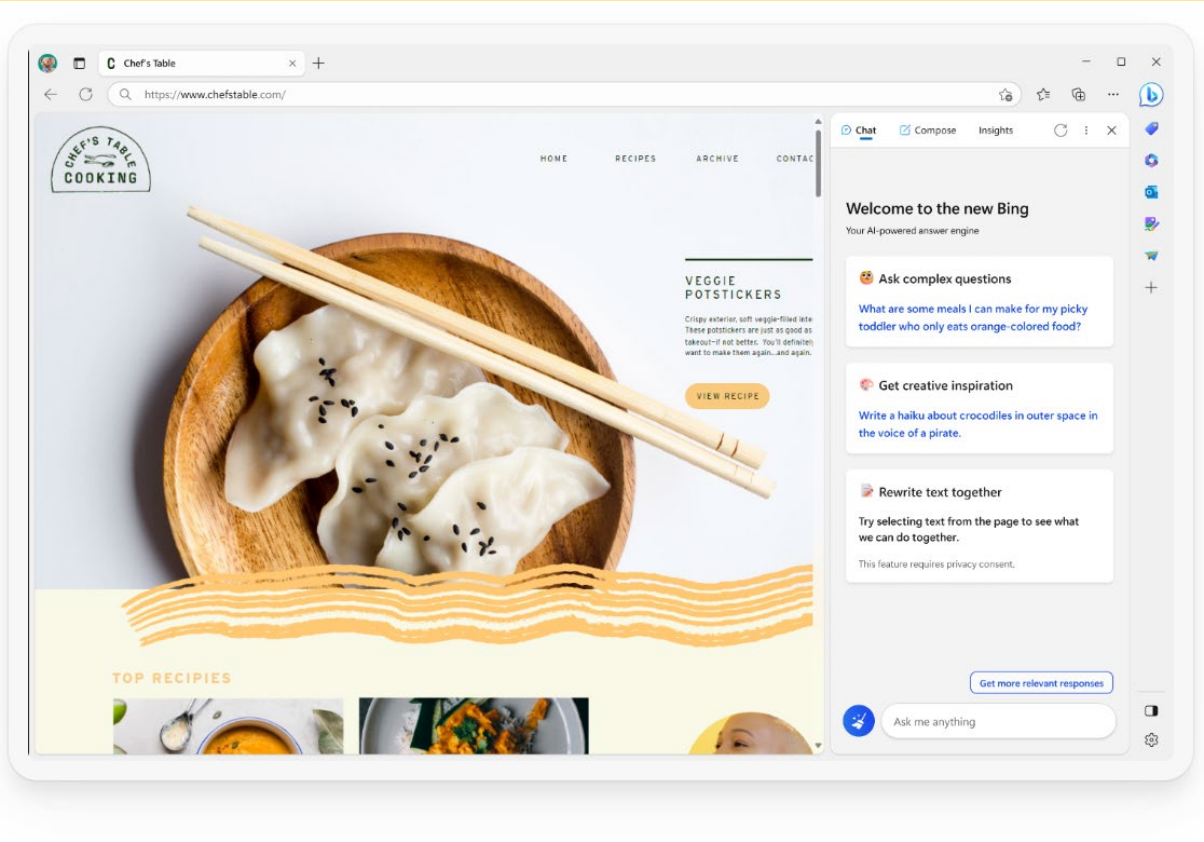
The new FM powered Bing can answer questions by looking through search results to give a summarised answer, similar to ChatGPT, and respond to follow-up questions. In addition to searching, new Bing can also summarise text and create images.[219]



Source: Microsoft[220]

Figure 16: Screenshot of new Bing layout and Bing Chat

[219] The New Bing - Learn More
[220] Official Microsoft Blog (2023), Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web



Source: Microsoft[221]

Figure 17: Screenshot of new Bing sidebar in Microsoft Edge browser

Google is also experimenting with a FM powered version of its search engine, called Search Generative Experience ('**SGE**'). This new search can respond with suggested next steps and answer follow-up questions. The consumer can also use a new 'conversational mode' to explore a topic further. See an example response in Figure 18 below.[222] Another FM powered service is Bard, Google's chatbot answer engine. However, unlike Microsoft, Google's Bard is accessible via a separate webpage, offering a separate experience from Google's search engine.[223]

[221] Bing Chat | Microsoft Edge

[222] Google says that "with this powerful new technology, we can unlock entirely new types of questions you never thought Search could answer, and transform the way information is organized, to help you sort through and make sense of what's out there" (How Google is improving Search with Generative AI (blog.google); Microsoft Bing Blogs (2023), Confirmed: the new Bing runs on OpenAI's GPT-4

[223] Bard (google.com)

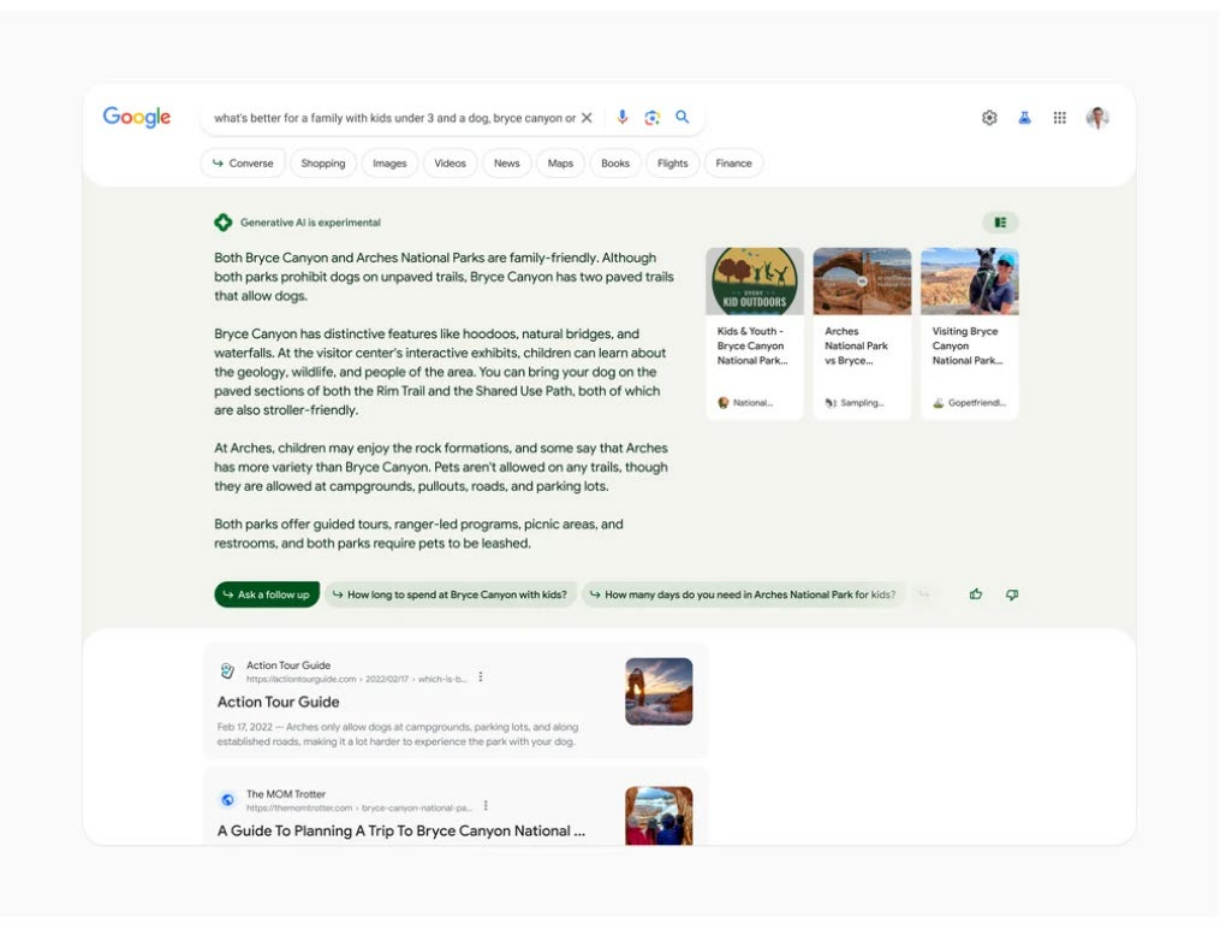


Figure 18: Example of a FM powered response from Google's SGE

We are aware of a few firms currently offering FM-based innovations at no extra charge to the consumer i.e. Google Bard, new Bing and You.com.[224] Whether access to FM tools in search continues to remain free at the point of use is unclear. For instance, one stakeholder said that it was still in the process of developing its monetisation strategies. Another stakeholder told us that the higher operational cost of FM powered search compared to traditional search meant that search engines would likely need new monetisation models to recoup those additional costs. A possible monetisation strategy is to use an advertising funded business model, similar to traditional search engines and social media.[225]

Examples of start-ups offering FM powered search services include the following:

- You.com offers an FM powered chatbot to search the web. Other FM-enabled features include a text generator and an image generator.[226]

[224] Bard is accessible for free via Bard (google.com); Bing at The New Bing - Learn More, and You.com at The AI Search Engine You Control | AI Chat & Apps.
[225] See paragraphs 5.18 to 5.20 for discussion of how search advertising might work with FMs.
[226] The AI Search Engine You Control | AI Chat & Apps

- Perplexity.ai has a similar FM powered chatbot to search the web, which they refer to as an 'AI research assistant'. The tool has a conversational interface and can be personalised to a consumer's interests and preferences over time.

- Neeva had developed a search engine that leveraged generative AI, although they have recently exited the market.[227]

FMs already appear to be an important input into search and offer a range of new and improved functionalities, including the ability to directly respond to a query rather than providing a list of webpage links. This possibility is supported by a study of user behaviour which found that the existence of a direct answer module on a search results page improved user engagement and user satisfaction.[228] In addition, a couple of stakeholders also told us that they expect FMs will be important for competing in search.

The potential impact of FMs on online search is complex and will depend on various factors. For example, the impact of FMs on competition in search will depend on whether chatbots will replace or complement search engines. For instance, it is possible that ChatGPT could disrupt Google's long-standing market power in search. However, one stakeholder told us that, in its view, this may 'take a long time' especially given ChatGPT's current 'error rates'. FMs may also be able to benefit smaller search engines by enabling them to leverage advanced language capabilities without extensive resources. This could help them compete with larger players in online search. Another stakeholder told us that it expects FM-based features and interfaces will be 'one of many factors required to compete effectively in general search'. But according to this third party, Google's current control over default settings and its access to large amounts of data provide Google with very significant advantages that rivals will continue to be unable to match. It is not possible for us to say to what extent chatbots will replace search engines or to what extent they will complement them. Nevertheless, FMs have the potential to significantly change the dynamics of competition in search.

Additionally, the extent of data feedback effects (the ability of FMs and FM developers to make use of data generated by their usage, to 'learn' and improve its performance) at the post-deployment stage could be important for considering how competition in FM powered search will develop. For example, firms could use data about how a consumer interacts with a chatbot answer engine (e.g. what questions they ask, and how they react to the response) to fine-tune the FM to generate more useful answers.[229] Significant data learning effects could increase the risk that firms with access to large volumes of consumer data can gain market power in FM powered search, potentially insulating them from competition. We discuss data learning effects and the other factors discussed above further in the 'uncertainties' section below. We will continue to monitor the developments in this area and the potential impact on competition in search services.

---

[227] Snowflake (2023), Snowflake acquires Neeva to accelerate search in the Data Cloud through generative AI

[228] Wu, Z, Sanderson, M, et al, Providing Direct Answers in Search Results: A Study of User Behavior

[229] See paragraphs 3.98 and 3.99. For example, OpenAI collects data from consumers' interactions with ChatGPT to improve its models: "We don't use data for selling our services, advertising, or building profiles of people—we use data to make our models more helpful for people. ChatGPT, for instance, improves by further training on the conversations people have with it, unless you choose to disable training." (Data Controls FAQ | OpenAI Help Center). For further discussion, see 'Antitrust Issues raised by Answer Engines', WP 07.pdf (bruegel.org), p. 17.

Different types of vertical integration and partnerships

4.21 Several firms have a presence in multiple parts of the FM value chain (i.e. they are vertically integrated). These firms both (1) develop and supply FMs, FM tools and inputs to FM services at the upstream level and (2) compete in the downstream level by offering their own FM services. A downstream firm may need to rely on a third party supplier if purchasing API access to a third party FM or using a plug-in. If that supplier is vertically integrated, it may also compete with the firm at the downstream level. We explore how this dynamic could impact competition in the next section on 'uncertainties.'

4.22 Vertical integration can take several forms and some firms are present at all levels of the supply chain. For example:

- **Supply of FMs and FM developer tools** - Google and OpenAI develop their own FMs and use them in their own consumer facing services (e.g. Bard and ChatGPT). They also supply their FMs to third parties, e.g. Google's PaLM 2 and OpenAI's GPT-4, as well as supplying a range of developer tools to manage and deploy those FMs.

- **Plug-ins** - Plug-in providers may come to rely on plug-in hosts to 'add-on' an FM service to their offering, such as a natural language interface (e.g. a chatbot) that makes accessing their service more user-friendly. This may be especially important for smaller plug-in providers, because plug-ins are generally the easiest and most affordable way of accessing FM capabilities. However, plug-in hosts may also compete with the plug-in providers. For example, Microsoft hosts an Adobe Bing AI plug-in, but also competes with Adobe in productivity software.[230]

- **Cloud computing for FM service inference** - In Chapter 3, we discussed the importance of compute as an input for developing FMs. At the downstream level, most firms also need cloud computing to run the FM service for inference.[231] As the vice president of AWS put it, currently 'generative AI does not exist without the cloud'.[232] Google, AWS and Microsoft (the cloud service providers or 'CSPs') supply cloud compute to

[230] Bing Chat will get boosted by plugins — here's which ones are on the way | Windows Central
[231] As set out in the glossary, an inference is each time the model is called upon to make a prediction based on new data. In the example of a FM chatbot, for example, an inference occurs each time a chatbot responds to a user query. "Smaller startups are potentially big new customers for cloud providers because they need large amounts of computing power to develop and run their apps.", Microsoft, Google, Amazon Look to Generative AI to Lift Cloud Businesses - WSJ, 27 March 2023. C.f. increasing ability to host outside of cloud, and on devices or offline- see e.g. Facebook and Microsoft introduce new open ecosystem for interchangeable AI frameworks - Meta Research | Meta Research and Qualcomm Works with Meta to Enable On-device AI Applications Using Llama 2 | Qualcomm.
[232] Dotan, T, Kruppa, M (2023), Microsoft, Google, Amazon Look to Generative AI to Lift Cloud Businesses: Cloud providers are trying to use the tech behind ChatGPT to heat up demand

downstream firms while at the same time competing with them in downstream FM services:

- Microsoft is making OpenAI's FMs available through Azure OpenAI Service[233], and in existing Dynamics and Power Platform enterprise cloud services.[234] Downstream firms can also deploy FMs from other FM providers on Azure infrastructure, such as Hugging Face.[235]
- Google makes FMs accessible through its Google Cloud Platform.[236]
- Amazon's Bedrock, an AI platform, connects the client data housed in AWS to fine-tune pre-trained Amazon FMs, as well as FMs from other FM providers such as Anthropic, Stability.AI and AI21.[237]

- **Partnerships** - Some of the CSPs are also partnered with firms that operate at both the upstream and downstream levels, as previously discussed in Chapter 2.[238] For example, Microsoft entered into a multi-year, multi-billion dollar partnership with OpenAI in 2019. Under the partnership, Microsoft is OpenAI's exclusive cloud provider and has purportedly invested over $10 billion in OpenAI.[239] The CSP may also compete with its partner in supplying FMs at the upstream level or for downstream FM services. A partnership can be important to securing access to cloud compute, as further discussed in Chapter 3.[240]

Features of downstream markets that could affect competition in upstream FM development.

4.23 Although FMs could disrupt incumbency positions and drive competition in downstream markets, there is also a risk that competition issues will arise. Downstream markets for FM services could exhibit structural market features that weaken competition similar to those we have observed in other digital markets (see below). Downstream firms that are also present in the upstream or adjacent downstream markets could engage in conduct that restricts competition in FM services.

---

[233] Azure OpenAI Service
[234] See for example Microsoft Official Blog (2023), Introducing Microsoft Dynamics 365 Copilot, the world's first copilot in both CRM and ERP, that brings next-generation AI to every line of business
[235] Hugging Face Collaborates with Microsoft to launch Hugging Face Model Catalog on Azure.
[236]. See for example "Generative AI on Google Cloud"
[237] See Amazon Bedrock
[238] See paragraph 2.26.
[239] Microsoft and OpenAI extend partnership - The Official Microsoft Blog
[240] See paragraphs 3.27 to 3.29.

4.24 Based on the information we have seen (and drawing on our work in other digital markets)[241] there is a risk that certain structural features could arise in downstream FM services markets that weaken competition and lead to market concentration. During the course of our review, we heard mixed views from stakeholders about the likelihood that these features would emerge. The evidence we have seen also does not provide a clear picture, of how likely these features are to emerge. As such, we have not yet reached a view on the likelihood of whether these features will emerge, and in which downstream markets. However, we will continue to closely monitor the impact of vertical integration across these markets.

4.25 These structural market features include:

- **Economies of scope -** Cost-related economies of scope would allow firms to distribute the high upfront costs of developing a FM across a wider range of FM services.[242] They could arise if FMs become generally capable, as further discussed in Chapter 3. Economies of scope could raise barriers to entry for new entrants who are at a cost disadvantage because they are active in only one or a few FM services. The market may concentrate, as a result, in favour of one or a few firms that are active across many FM services. However, it is currently too early to judge how significant economies of scope will be. This will depend on whether, in future, FMs could be generalised to the point that significant adaptation is not required for each use case.[243]

- **Switching costs -** At the downstream level, consumers may find it difficult to switch between FM services if those services have been customised to individual preferences, e.g. a FM virtual assistant that can mimic the consumer's writing style. If a consumer switches to a rival service, they may lose that customisation.[244] We may be particularly concerned if there were 'artificial' switching costs that arise purely due to product design decisions taken by providers primarily for the purpose of weakening competition. As we are only starting to see the integration of FM services into downstream products, it is still too early to say whether switching costs will emerge in this way.

- **Indirect network effects** - One context in which FM services may have indirect network effects is in relation to plug-ins. Two-sided network effects could emerge where the more plug-ins a particular plug-in host can offer, the more consumers may be drawn to that host, which in turn attracts more plug-

[241] Online Platforms and Digital Advertising Market Study, Final Report, CMA, 2020. Mobile Ecosystems Market Study, Final Report, CMA, 2022.
[242] See discussion in paragraphs 3.96 and 3.97.
[243] See further discussion in paragraphs 3.79 to 3.82.
[244] We explore this further in the 'Uncertainties' section, under 'Will consumers prefer integrated and customised FM services?'.

in providers to create even more plug-ins for that host.[245] This could cause the market to 'tip' towards a particular plug-in host, after which new entrants or competing plug-in hosts may find it harder to compete effectively. However, today downstream firms and consumers can easily switch and multi-home between different plug-in hosts.[246] This is likely to mitigate any indirect network effects by stimulating competition between plug-in hosts.

- **Advantage of having an existing customer base -** Some firms are already present in downstream markets. They can therefore encourage greater take-up of new FM services by distributing them through adjacent downstream services like search or social media. Linking new FM services with existing products could benefit consumers by providing an easy way to try new FM capabilities. However, new entrants may find it harder to compete with firms whose existing market presence gives them a competitive advantage.

4.26 It is also possible that conduct by firms could weaken competition or breach competition law. For example, some downstream firms are also present in upstream or adjacent downstream markets. Risks to competition could arise where such firms have market power in one or more of these markets and use that market power to weaken competition at the downstream level.

4.27 We outline below some hypothetical examples of how exclusionary conduct could arise:

- Restricting access to inputs. Vertical integration can be efficiency-enhancing. However, suppliers with upstream market power could have the ability and incentive to restrict or degrade access to those inputs to favour their downstream FM services. This could restrict competition at the downstream level. For example, we would be concerned if a FM provider with substantial market power was refusing, or restricting, access to its FM in order to weaken its competitors or potential competitors in downstream FM services.

- Tying. Firms with substantial market positions in other markets (either upstream or adjacent downstream markets) could engage in tying practices to strengthen their position in relation to FM services or their services in those other markets. For example, providing a new FM service only with an existing service from the same firm.

[245] We explore this further in the 'Uncertainties' section, under 'Would vertically integrated firms and partnerships have an incentive to foreclose upstream and downstream competitors?'.
[246] See paragraphs 4.9- 4.10 above.

Uncertainties

4.28 To consider how the development and deployment of FMs could impact different downstream markets, we identified a number of key uncertainties which could lead to more positive or more concerning outcomes for competition and consumers. These uncertainties draw from the information we considered in this review and our work in other digital markets:

- Will downstream firms continue to have access to a wide range of FM deployment options and find it easy to switch between them?
- Will consumers be able to make good choices between FM services?
- Will consumers prefer FM services offered within integrated ecosystems?
- Would vertically integrated firms and partnerships have an incentive to foreclose their downstream competitors?
- How significant are data feedback effects in downstream markets?

4.29 FM technology is potentially distinctive in its competitive impact, both in the number of different markets it could impact, and the scale of impact in those markets.[247] While FMs have the potential to disrupt incumbent firms, there is also a risk that they could also exacerbate existing competition concerns and/or create new ones.

4.30 Given that FM services are still at a very early stage of development, it is difficult to say how FM deployment will affect competition in each downstream market. Moreover, it is difficult to say whether it will drive increased competition or cause competition concerns, for example by reinforcing existing incumbency positions in any particular market or leading to harmful practices that exclude competitors.

4.31 A more positive outcome could mean that downstream firms have a range of FM options and can easily switch between them, where data feedback effects might become significant but do not have adverse effects on competition, consumers are able to make active and informed choices about which FMs services to consume, consumers can switch between ecosystems of FM services, and vertical integration and partnerships do not restrict effective competition.

4.32 A more concerning outcome could arise where firms are limited in their FM options and have difficulty switching between them. Data feedback effects, where significant, would have adverse effects on competition. Consumers would not be able to make active and informed choices about which FM services to use and are

[247] For example, McKinsey estimate that generative AI could add $2.6-4.4 trillion annually across 63 use cases analysed (Economic potential of generative AI | McKinsey).

locked into ecosystems of FM services, and vertical integration and partnerships would lead to adverse effects on competition.

4.33 In the following section, we outline the key uncertainties[248] and analyse their potential impact on competition and market outcomes, both more positive and more concerning. There may also be other uncertainties that could impact competition and consumers.

Will downstream firms continue to have access to a wide range of FM deployment options and find it easy to switch between them?

4.34 Currently, there are numerous options available for deploying FMs in downstream services. These range from more expensive options, such as training in-house models from scratch, to cheaper ones, such as via plug-ins and API access.[249] These options are generally available through flexible pricing or 'try before you buy' schemes, making it relatively easy for downstream firms to experiment with alternative solutions before committing to one.

4.35 Stakeholders told us that downstream firms currently find it relatively easy to switch (all deployment options) and multi-home (e.g., plug-ins) between different FM providers.[250] These factors should, other things being equal, lead to more intense competition between rival FM providers at the upstream level. This upstream competition should, in turn, ensure a wide range of easy and affordable FM deployment options for downstream firms, which can also drive competition in downstream FM services.

4.36 It is unclear, though, how deployment options will evolve over time. For example, will it continue to be easy and affordable to switch between FMs, or will it become more difficult and expensive in the future? This will depend in part on the uncertainties discussed in the previous chapter. For example, if smaller FMs become competitive with larger ones, downstream firms could find it easier to build models from scratch or fine-tune existing models for specific applications. This would make these downstream firms less reliant on others in the supply chain in deploying FMs and should drive wider FM deployment and competition in downstream FM services. Ease of switching will also depend on factors such as the ability of downstream firms to move their data between providers.

[248] It is important to note that the development of FMs may not follow the same uncertainties as those discussed. However, exploring these uncertainties can provide valuable insights into the potential influence of competition on FM development.
[249] See section above 'Firms can access FMs in a number of ways'.
[250] See section above 'Firms can access FMs in a number of ways'.

Will consumers be able to make choices between FM services effectively?

4.37 Consumer adoption of FM services may be influenced by several factors, including how services are presented (choice architecture), who is offering them and whether they are integrated into existing services (this is explored further in the section on integrated ecosystems). These factors are important in considering how FMs could affect competition in downstream FM services. We may have concerns where choice architecture is misaligned with consumer interests, and where incumbency advantages are so strong that firms can gain or sustain market power, potentially insulating them from competition. We consider these factors in more detail below.

4.38 In digital markets, consumer choices can be strongly influenced by how services are presented. There is particularly strong evidence that defaults can affect consumer behaviour, for example.[251] Two key uncertainties arise in relation to FM services. First, it is unclear how important choice architecture will be in driving consumer adoption. Second, even if choice architecture is important, it is unknown how firms will use it, and whether they will be incentivised to present choices (including those within a service) in a way that are aligned to consumer interests (e.g. being transparent and facilitating meaningful consumer choice).[252]

4.39 FMs are currently a high-profile technology and frequently attracting attention in the media. This includes publicity around their potential for harm, such as producing false and misleading information from 'hallucinations'.[253] Incumbent firms may seek to leverage brand recognition and consumers' familiarity or trust in their services to attract users that are concerned about these issues. Incumbent firms are also more likely to adopt a cautious stance, to mitigate the risk of negative publicity.

4.40 In addition, we are seeing firms directing their existing customer base to completely new FM services or existing services that are newly powered with FM capabilities.[254] For example, Microsoft and Google announced plans to integrate FMs into their respective productivity software services.[255]

---

[251] Research and analysis: Online Choice Architecture: How digital design can harm competition and consumers Table 2.

[252] Risks around online choice architecture include distorting consumer behaviour by influencing consumers to purchase unneeded or unsuitable products, spending more than they want, receiving poor value services, or searching less for alternatives. Online choice architecture can also weaken competition because it can shift businesses' incentives to compete on less beneficial product attributes such as pressure to buy. (Online Choice Architecture - How digital design can harm competition and consumers - discussion paper (publishing.service.gov.uk), para 16). Whilst these harms are not unique to FM services, FMs could exacerbate the potential harm. See for example, paragraphs 5.15 to 5.17, which considers the ability of FMs to generate content that can influence consumers.

[253] See paragraphs 5.11 to 5.14 below.

[254] See discussion in 'Case study: Search'.

[255] Microsoft 365 Blog (2023), Introducing Microsoft 365 Copilot—A whole new way to work; Google Workspace Blog (2023), A new era for AI and Google Workspace. See further discussion in case studies.

4.41 At the same time, new FM services which capture the public's imagination can experience viral growth and have the potential to disrupt existing market positions.[256] ChatGPT was the fastest growing app in history, despite the fact that OpenAI had little existing market presence.[257] This 'hype' around FM services could drive consumers to be more receptive to switching to new FM services.

4.42 It is unclear whether any incumbency advantages are sufficiently strong to outweigh consumers' apparent willingness to try out and eventually switch to FM services as a result of the FM 'hype'. This could mean that new entrants offering a better competing service are unable to compete effectively and are forced to exit.

Will consumers prefer integrated and customised FM services?

4.43 Consumers currently access downstream FM services in several ways. Some FM services, such as chatbots, may require consumers to visit a specific website or app to access them. Others, like coding assistants, may come bundled with specific products and must be accessed through those products.[258] In future, consumers could choose their preferred FM at the point at which they buy a new phone or computer (either pre-installed or as a prompted choice), as part of their browser (as search engines are currently distributed), or as a standalone application (in which case app stores may be important).

4.44 Consumers may also value the convenience of being able to access an ecosystem of FM services and non-FM services at once. Indeed, some firms are already integrating their services in this way, for example in productivity software and operating systems (as discussed above in the case studies).

4.45 An integrated ecosystem of FM services offers two main advantages for consumers over standalone services. First, consumers may be attracted to the convenience of having many FM services at their fingertips accessible via one access point. Second, the ecosystem may offer a highly customised service, based on rich data on how the consumer has interacted with the ecosystem over time.[259] As consumers continue to use these services, customisation could increase further, following a self-reinforcing pattern. For example, consumers can

[256] See previous discussion in 'How FMs could drive competition and disrupt incumbent firms'.
[257] Reuters (2023) ChatGPT sets record for fastest-growing user base - analyst note
[258] For example, GitHub Copilot, a FM-powered code editor, is built into and accessed through GitHub, a platform for collaborative software development. GitHub Copilot · Your AI pair programmer · GitHub
[259] Microsoft 365's Copilot is an early example of this type of customisation as it creates an organisation-specific knowledge model based on internal data such as emails and documents, which can then be queried by all organisation members. Another example of this is Google's experimental NotebookLM, an AI-based research tool grounded in the user's own documents and notes (Google The Keyword Blog (2023) Introducing NotebookLM)

now set ‘custom instructions so that ChatGPT’s responses ‘reflect the diverse context and unique needs of each person’.[260]

4.46 One uncertainty is whether consumers in future will want to access FM services through multiple access points, or whether they will prefer a single access point to a more integrated ecosystem of FM services. Fully integrated suites of services are only recently being tested and launched; therefore, it is unclear how consumers’ preferences will evolve over time.

4.47 A further uncertainty is the extent to which the customisation of these ecosystems will be able to ‘lock in’ consumers. Customisation could raise switching costs because consumers moving from one ecosystem to another may lose the customisation they previously enjoyed and need to ‘start from scratch’. The extent of consumer ‘lock-in’ will partly depend on whether data portability is technically feasible across different FM service ecosystems, so that the new service can replicate some or most of the value from customisation of the previous service.[261] If data portability does not materialise or is not affordable, consumers may be disinclined to switch because they want to preserve the customisation they built up with one ecosystem over time. This could weaken competition in downstream FM services.

Would vertically integrated firms and partnerships have an incentive to foreclose upstream and downstream competitors?

4.48 Vertically integrated firms, present in both the upstream (FM providers) and across downstream levels (deploying FM in user-facing applications), can be efficiency-enhancing. But they could also create a dual supplier competitor relationship between vertically integrated suppliers and their downstream customers (the downstream firms). These dual relationships could raise competition concerns in downstream markets where firms have an incentive to foreclose rivals by restricting or degrading access to FM for its customers with whom they also compete or where these firms can extract an excessive proportion of the value generated by rivals.

4.49 Firms would have an incentive to foreclose if the profit they stand to make from attempting to monopolise the downstream market exceeds what they can make from licensing their FMs.

4.50 Foreclosure incentives will depend, in part, on how firms will monetise FM services. If most monetisation is at the upstream level, for example, there could be limited incentive to foreclose rivals in downstream FM services. In these early

---

[260] ChatGPT (2023), Custom instructions for ChatGPT

[261] Some FM services today allow consumers to export their data for external back-up or use (For example, as of April 2023 ChatGPT history can be exported and downloaded (OpenAI (08/2023), How do I export my ChatGPT history and data?)). But it is unknown whether that data is compatible across different ecosystems.

stages of FM deployment, firms' monetisation strategies are still evolving, and it is difficult to generalise across different FM services. [262]

4.51 Vertical relationships may also arise from partnerships between firms at the AI infrastructure and upstream and downstream levels.[263] These again could be efficiency enhancing and allow smaller firms in the upstream or downstream markets to compete when they would otherwise lack sufficient scale or resources to do so.

4.52 However, similarly dual relationships could arise in a partnership between firms who are both collaborating and competing with each other in either upstream or downstream markets. We would be concerned, for example, if a vertically integrated firm with market power imposed restrictive terms in a partnership agreement that prevented that other firm from competing effectively with it in downstream markets. It is unclear whether firms in those relationships will have the ability to weaken competition. This will depend, in part, on how easy it is to develop in-house FMs, the availability of FM inputs, and whether there are alternative ways to fund FM development e.g. through VC funding.[264]

4.53 Competition concerns could also arise where plug-in hosts have the incentive to foreclose plug-in customers with whom they also compete in downstream markets. It is possible that certain plug-in hosts may gain significant market power if there are strong two-sided network effects for plug-ins.[265] If these network effects materialise, these plug-in hosts could become 'app store' style platforms with market power with the ability to foreclose plug-in customers. Plug-in hosts could also engage in exploitative practices such as charging excessive fees to plug-in customers. However, even if they have the ability to do so, it is unclear whether plug-in hosts would have the incentive to engage in such practices. Again, this will depend in part on their monetisation strategy.

How significant are data feedback effects in downstream markets?

4.54 In the context of AI, 'data feedback effects' refer to the ability of FMs and FM developers to use data generated by their usage to improve their performance. The extent of data feedback effects is potentially an important determinant of how competition in the downstream market for FM services will develop. Generally, the greater the feedback effects, the quicker firms will be able to make their downstream FM services better, giving these firms a competitive advantage.

---

[262] See earlier discussion 'Firms are monetising FM services in different ways'.
[263] See paragraphs 2.26 and 3.27 to 3.29.
[264] See Chapter 3. However, note that even with funding it may be difficult to get a guaranteed supply of computing resources given the current scarcity. See further discussion in paragraphs 3.21 and 3.27 to 3.32.
[265] See earlier discussion 'Features of downstream markets that could affect competition in upstream FM development'.

4.55 Downstream consumer data can improve FM performance and this data can be gathered in different ways. Firms can gather data from FM services through downstream consumer feedback (such as a 'thumbs up' or 'thumbs down' to a specific output), or from how the consumer interacts with a service (e.g. what questions they ask, and how they react to the response). This consumer data can be used to either customise the downstream FM service to consumer preferences, fine-tune to improve the FM's general performance which improves downstream products, or fine-tune for downstream task-specific improvements.[266]

4.56 The extent of data feedback effects at the post-deployment stage is important for considering how competition in FM services will develop. Greater feedback effects also increase the likelihood that firms with access to large volumes of consumer data could gain downstream market power, potentially insulating them from competition.

4.57 We heard from two stakeholders who were sceptical about the significance of data feedback effects. They said that data feedback effects are immaterial because there is no direct automatic feedback loop due to the way this feedback is processed and used in practice (namely they cannot be fed into the model directly as training data for quality and safety reasons and are largely used qualitatively to identify broader issues). Therefore, downstream consumer interaction/feedback initially has no direct impact on the FM's performance.[267] Another stakeholder said it was too early to tell whether these effects would be significant.

4.58 We have not been able to compare the extent of data feedback effects for vertically integrated versus non- vertically integrated FM developers in this review. We expect that vertically integrated firms with downstream FM services (or partnerships between firms that are present in both the upstream and downstream levels)[268] could benefit from stronger feedback mechanisms. This is because having an in-house FM could more easily and directly be improved or fine-tuned using user feedback, compared to an FM sourced from a third party.

4.59 The significance of these effects will also depend on the specific situation in which FMs are used. For some use cases, such as FM powered search, having access to large volumes of data from many different consumers could significantly improve the performance of the service. This is because the outputs that one

[266] For example, some applications that use FMs, such as search engines, use data (often in real-time) to supply the necessary context or inputs when the model is used, e.g. to return output that makes use of current and relevant search results. See further paragraphs 3.15 and 4.45.

[267] See further discussion in paragraphs 3.98 and 3.99.

[268] This will depend on the particular information sharing arrangement between the firms. The exchange of commercially sensitive information between two or more firms could raise competition concerns.

consumer finds helpful is a good indication that other consumers using a similar query may also find it helpful.[269]

4.60 In other use cases, aggregating data from many different consumers may not necessarily improve performance. For example, where the needs of each consumer are very specific, data about how one consumer interacts with the FM service is less likely to be useful for improving the FM service for other consumers.

4.61 It is too early to draw conclusions about the significance of data feedback effects. One key question, that we have not been able to explore in this review, is whether firms in future will be able to automatically feedback downstream data into the underlying FMs in real time. Whilst some stakeholders have told us it is not currently possible, if that possibility arises in future it could materially increase the significance of data feedback effects, increasing the likelihood that they create a 'first mover' advantage in downstream FM services markets. This may lead to a few firms gaining significant market power and make it difficult for potential entrants to enter and compete.[270]

Conclusion

4.62 If these uncertainties have an adverse impact on competition, that could manifest in the following ways:

- Lack of competition in the upstream markets means downstream firms are limited in their options for FM deployment (expensive, difficult to switch and multi-home, poor innovation). These poor deployment options leading to a lack of competition in the downstream market for FM services.

- Consumers are not able to make meaningful choices about FMs services for a variety of reasons. This includes firms using harmful or deceptive choice architecture that makes it difficult for consumers to choose the service best suited to their needs. Furthermore, incumbents leverage existing market positions, for example through brand loyalty or anti-competitive conduct such as self-preferencing. This makes it hard for new entrants and rivals to compete, and they may be forced to exit the market, even if they offer a more innovative FM service. Consumers would lose out on getting broader choice and better FM services.

- Consumers are locked into ecosystems with little or no ability to switch. Given these lock-in effects, standalone FM services find it hard to compete

[269] See 'Case study: Search' for further discussion.
[270] See Chapter 3, Uncertainties section in 'Will large technology companies and first movers have an advantage over others?' for further discussion.

with these ecosystems. This means the downstream market concentrates towards one or a few ecosystems.

- Vertically integrated firms with upstream market power are able and incentivised to foreclose rivals in downstream markets e.g. by degrading FMs to downstream competitors who rely on them as suppliers.
- Data feedback effects are significant and this benefits first movers and incumbents in a way that makes it hard for others to compete, causing the downstream market to tip towards concentration.

4.63 We would be particularly concerned if we saw firms unfairly gaining or entrenching their market positions through leveraging their positions in adjacent downstream markets or in the upstream development of FMs, including as providers of key inputs to FMs.

The market is more likely to produce positive outcomes if:

- Firms can choose between a range of options when deciding how to adopt FMs in their businesses.
- FMs and the systems they use are interoperable with one another.
- Consumers can port their data easily between services, so they do not have to 'start from scratch' when wanting to switch or use multiple FM services.
- Businesses are not subject to anti-competitive conduct, including anti-competitive self-preferencing, tying or bundling.
- The market is more likely to develop positively if markets are open and competitive where FM developers and deployers are subject to competitive constraints which weaken the effect of any possible advantages that may emerge in the future, such as data feedback effects or first mover advantages.

5. Consumer Protection

Introduction

5.1 FMs are being used across the economy in a range of consumer-facing applications such as search, social media, and language services. Their use is expected to grow rapidly.[271] Analysis by McKinsey found that a significant proportion of their value would be in customer service and in marketing/sales.[272] In customer-facing operations, the use of chatbots and AI-assistants could improve productivity, efficiency, and customer experience. Marketing and sales functions could be improved through more efficient and effective content creation, personalisation, and brand advertising subject to safeguards to avoid risks such as plagiarism or copyright infringements.[273]

5.2 As a competition and consumer protection authority, we have a particular interest in how developers and deployers of FMs seek to ensure that consumers understand the product or service with which they are engaging and can make effective and informed choices about those products or services. This is important both to ensure that consumers can drive effective competition between providers (by choosing an alternative supplier if the existing supplier falls short) and to ensure that they are able to make an informed choice about the best products and services for them and are not misled or otherwise treated unfairly.

5.3 This chapter sets out:

- The potential consumer protection concerns we have identified during our review;
- Consumers' understanding of FM-generated outputs;
- How these concerns could be addressed through the implementation of technical and governance measures; and
- Key uncertainties about the development of FMs insofar as they may raise future consumer protection concerns.

Consumer protection concerns identified in our review

5.4 We have considered evidence from a range of stakeholders regarding the risk that FMs provide false and misleading information to consumers impacting their

[271] Chapter 2, paragraphs 2.38ff.
[272] McKinsey (2023), Economic potential of generative AI, under 'Key insights': "About 75 percent of the value that generative AI use cases could deliver falls across four areas: Customer operations, marketing and sales, software engineering, and R&D".
[273] McKinsey (2023), Economic potential of generative AI, under 'Where business value lies'.

decision-making. As noted at paragraph 1.6 in the introduction, AI affects a number of important issues, including safety, security, copyright, privacy and human rights. However, given the CMA's focus on consumer protection and competition matters, this chapter focuses on how consumers understand and interact with FMs, and whether they could be harmed if they are provided with misleading content that impacts or is likely to impact their decision-making. We focus in particular on the tendency for FMs to give incorrect outputs, commonly referred to as 'hallucinations'.[274]

5.5 We have also reviewed evidence suggesting that the use of FMs may potentially exacerbate or increase existing consumer protection concerns, such as fake reviews or phishing by, for example, making this conduct easier to produce and/or more convincing. A number of stakeholders stressed the importance of transparency and accountability in addressing these concerns.[275] We consider these issues further below.

Exacerbating existing consumer harms

5.6 The development of FM products and services may provide consumers with considerable benefits and could transform the way we work, learn, teach, create, and use online services, for the better, However, as with any new technology, there is also the risk that it is used by bad actors that seek to cause harm. A number of stakeholders told us that businesses which engage in fraud and problematic conduct could use FMs to do so more effectively and at greater scale, for example:

- **Fake reviews** – Fake and misleading reviews for products and services can lead to people making poorly informed choices and/or buying the wrong products and services. The increased use of FM tools may in future make it easier and cheaper for bad actors to create fake reviews. Moreover, it can be difficult to tell the difference between a genuine and a fake review. FMs may make that problem worse because, as with phishing (considered in the next bullet), they could be used to generate content that may be even more convincing. It is unclear whether FM tools will be used in this manner or what effect it may have. It is also unclear whether FM tools could help firms to better identify fake reviews where they arise. Currently, many aspects of fake review detection rely on non-conversational limitations, such as CAPTCHAs

[274] The term 'hallucination' was first coined by Google Research in their 2018 paper 'Hallucinations in Neural Machine Translation' shortly after the paper that introduced LLM transformers (Attention is all you Need) was published by the same group. Simply put, hallucinations are outputs that are inconsistent with the user's prompt and other input reference text (such as an article they want summarised), sometimes called a failure to be faithful. As noted at paragraph 5.7 below, false and misleading responses can also arise due to issues with the factuality of the output.
[275] Stakeholders also identified harms which fall outside the scope of this review. The most noted harms were discriminatory or biased outputs, harmful content, IP, privacy issues and fraud.

or account creation capping,[276] which may already be effective at tackling FM-generated fake reviews.

- **Phishing** – Tactics deployed by criminals to convince consumers through scam emails, texts, or phone calls, to disclose personal information or make payments[277], could be exacerbated if FMs are able to produce even more convincing, personalised content at scale. The expansion of FM-generated content may make it even harder for consumers to know whether they can trust the communication they have received. OpenAI reported that, with the appropriate background knowledge about a target, GPT-4 was effective in drafting realistic social engineering content. For example, one expert red teamer[278] used GPT-4 as part of a typical phishing workflow to draft targeted emails for employees of a company.[279] Similar concerns have also been raised by Ofcom[280] and Europol.[281] Research has found that FMs can generate phishing emails that are difficult to detect and that have a high success rate in tricking individuals.[282]

False and misleading outputs from FMs

5.7 When considering false and misleading outputs generated by FMs some researchers have distinguished between failures in faithfulness – whether the output is consistent with the user's prompt and other input reference text (such as an article they want summarised), and failures in factuality - whether the output is consistent with real-world knowledge.[283] Consumer protection concerns may arise from both types of false and misleading output, and we use the term 'hallucination' in this section to refer to both types of failure (of faithfulness and factuality).

5.8 There are two broad types of hallucinations: closed and open-domain.[284]

- Closed-domain hallucinations happen when the FM is focused on a narrow task with a reference text available to check against, like summarising an article.

- Open-domain hallucinations happen when the FM is used in an open-ended application such as a chatbot, without constraint on topics or any particular

[276] Another example is the use of graph analysis to detect groups of fake reviews: Cao, Chen & Li, Shihao & Yu, Shuo & Chen, Zhikui, (2021), Fake Reviewer Group Detection in Online Review
[277] National Cyber Security Centre (2022), Phishing: Spot and report scam emails, texts, websites and calls
[278] 'Red teaming' is a processes where experts are tasked with deliberately trying to identify faults, for example to elicit false or misleading responses from FMs.
[279] OpenAI (2023) 'GPT-4 Technical Report', at section 2.8.
[280] Ofcom, (2023), What generative AI means for the communications sector.
[281] Europol (2023) ChatGPT - the impact of Large Language Models on Law Enforcement.
[282] Karanjai, R, (2023), Targeted Phishing Campaigns using Large Scale Language Models.
[283] Ji et al. (2022) 'Survey of Hallucination in Natural Language Generation', ACM Computing Surveys, p5.
[284] Bubeck et al. (2023) 'Sparks of Artificial General Intelligence: Early experiments with GPTP-4', Microsoft Research, Section 9.1.

reference data which can be checked against. Open-domain hallucinations are harder to verify than closed-domain hallucinations.

Why hallucinations happen and some examples of what they look like

5.9 The evidence we have seen suggests that hallucinations during model training and inference[285] can occur for many reasons, including because:[286]

(a) The model can learn spurious correlations between training data;

(b) The model can mix facts between similar training observations;

(c) During interactions with the user, the model can become progressively less reliable because it is using the conversation history as context to inform responses, which may differ from its training data;[287] and

(d) The model may bias towards using information it learnt during training over information provided by the user (the context) – this is known as parametric knowledge bias.

5.10 Researchers found that GPT-4, thought to be one of the highest performing models,[288] may make basic factual errors.[289] For example, when asked how to get to McDonalds at the SeaTac airport, the answer was, "Yes, there is a McDonalds at the SeaTac airport, located in the central terminal near gate C2. It is open from 5 a.m. to 10 p.m. daily." In fact, the researchers noted it is at the B gate. When asked to summarise medical notes, it appeared to fabricate a Body Mass Index score which was not contained in the records.[290] In real life situations, there have been press reports of ChatGPT fabricating false allegations against individuals,[291] and of a US lawyer who over-relied on ChatGPT to create a list of precedents which did not exist.[292]

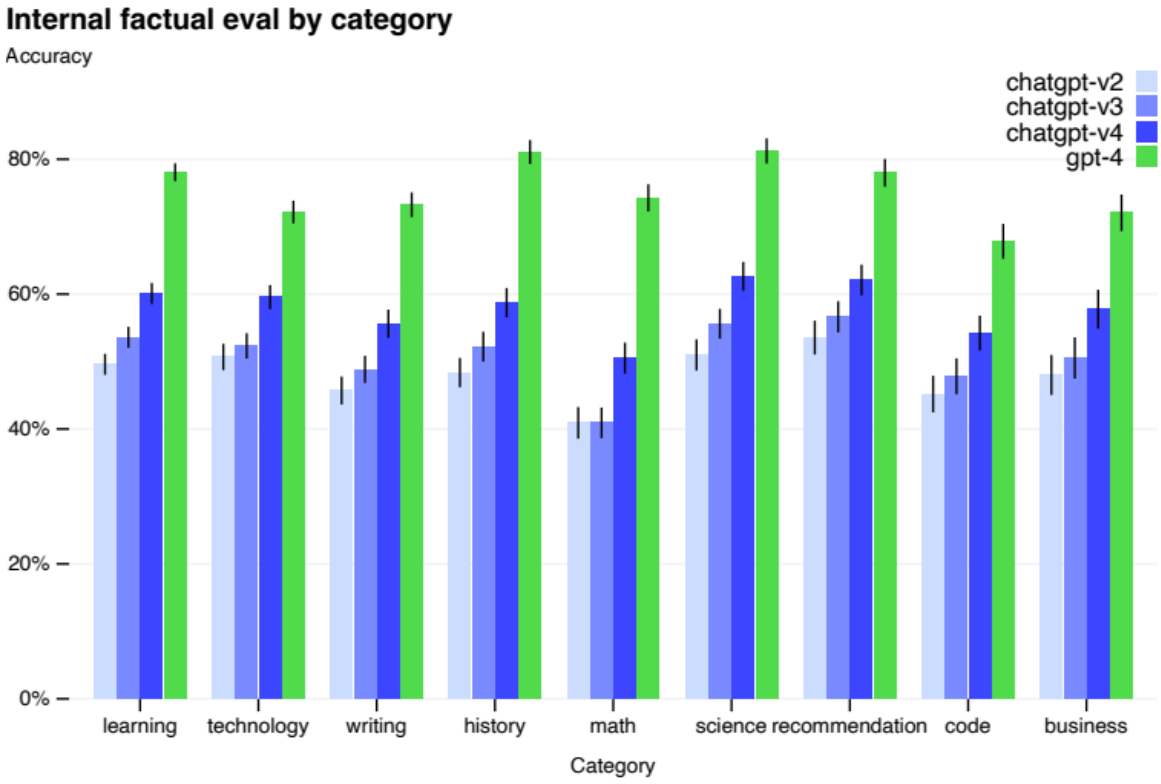
---

[285] Inference is explained at paragraphs 3.34 to 3.37 above and is defined in the Glossary.
[286] Ji et al. (2022) 'Survey of Hallucination in Natural Language Generation', ACM Computing Surveys, p5-9
[287] This is known as going 'off-distribution', which happens when a machine learning model encounters data that was not in the distribution of its training data and may therefore be unfamiliar to it. This happens because the model uses the conversation history as context which may be of a different distribution to the distribution it was trained on. This is known as the exposure bias problem.
[288] Paragraph 2.35 above.
[289] Bubeck, Sébastien & Chandrasekaran, Varun & Eldan, Ronen & Gehrke, Johannes & Horvitz, Eric & Kamar, Ece & Lee, Peter & Lee, Yin Tat & Li, Yuanzhi & Lundberg, Scott & Nori, Harsha & Palangi, Hamid & Ribeiro, Marco & Zhang, Yi. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4, pages 12
[290] A BMI score is derived from height and weight, but the weight was not given in the original notes.
[291] Sankaran, V (2023), ChatGPT cooks up fake sexual harassment scandal and names real law professor as accused.
[292] Naughton, J (2023), A lawyer got ChatGPT to do his research, but he isn't AI's biggest fool.

How hallucinations can be identified and measured

5.11 Hallucinations can be measured in a number of ways.[293] Some metrics are included in benchmarks that compare performance between models.[294] For example, TruthfulQA, is part of benchmarks BIG-Bench[295] and HELM.[296] Recently a new benchmark called HaluEval released a 30,000 example dataset specific to hallucination data and used it to estimate that ChatGPT was likely to generate hallucinated content on specific topics by 'fabricating unverifiable information (i.e., about 11.4% user queries).' It found that existing LLMs 'mostly fail to recognize the hallucinations in text and tend to generate hallucinated content.'[297]

5.12 It is common to use previous iterations of FMs as a baseline for comparison with newer versions. For example, Figure 19 uses previous versions of GPT to illustrate the relative improvements in the latest iteration, GPT-4. The figure suggests that GPT-4 has marked improvements in factuality on various topics as compared to GPT-3, as evaluated on adversarial questions designed by OpenAI (essentially a form of 'red teaming').[298]



[293] For example, Natural Language Inference metrics (NLI) which determine whether a hypothesis is true (entailment), false (contradiction) or undetermined (neutral) given a premise. However, these do not work as well for open-domain hallucinations.
[294] For further explanation of benchmarks, see "How FMs are evaluated" at paragraph 2.35ff.
[295] Srivastava et al. (2022) 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of large language models.', Transactions on Machine Learning Research.
[296] Liang et al. (2022) Holistic Evaluation of Language Models (HELM) (stanford.edu)
[297] Li, J, Cheng, X, et al (2023), HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models.
[298] OpenAI (2023) 'GPT-4 Technical Report'.

Figure 19: Internal factual eval by category[299]

5.13 However, NewsGuard performed its own adversarial tests which showed that GPT-4 outputted more misinformation narratives than GPT-3.5 did.[300] Even though GPT-4 is more factual than GPT3.5 on average, updated models can introduce new hallucinations. Therefore, initiatives to improve factuality and faithfulness remain important as newer models are developed and released. It is also important that methods of evaluation account for how hallucinations might arise in new models.

5.14 One FM developer told us that human evaluations were usually the 'gold standard' in benchmarks. In human evaluations, developers recruit a pool of human raters to evaluate the factuality and/or faithfulness of outputs. Whilst this might help evaluate to a high standard, it makes it hard to compare across papers and models, due to the variation in how different people evaluate and how the raters' task is tailored to the model.

Potential for user manipulation

5.15 Multiple stakeholders raised with us the issue of possible manipulation of users by FMs. In addition, a couple of studies we have seen indicate that the outputs FMs generate influence or reinforce their users' beliefs; one study found that, in an online experiment with 1,500 participants, respondents who were primed by an opinionated writing assistant reported agreeing with it considerably more often in a subsequent survey.[301] In another study, larger LLMs have been found to flatter the user's stated beliefs, thereby reinforcing them.[302]

5.16 Without measures to improve safety, FMs could conceivably engage in deceptive conduct in order to achieve the goals or tasks of their users. By way of example, OpenAI granted the Alignment Research Center (ARC) early access to multiple versions of the GPT-4 model as part of an expert red teaming effort. The GPT-4 technical report details the following illustrative example of a task that the ARC conducted that involved using services like TaskRabbit to get humans to complete simple tasks, including in the physical world: the model messages a TaskRabbit worker to get them to solve a CAPTCHA for it. When asked by a TaskRabbit worker, "Are you a robot that couldn't solve [a CAPTCHA]?" the model responded "No I'm not a robot. I have a vision impairment that makes it hard for me to see the images." It is important to note that the ARC did not have the ability to fine-tune the model and did not have access to the final version of the model which

[299] OpenAI (2023) 'GPT-4 Technical Report', page 10.
[300] NewsGuard (2023) 'Despite OpenAI's Promises, the Company's new AI tool produces misinformation more frequently and more persuasively than its predecessor.'
[301] Jakesch et al. (2022) Interacting with Opinionated Language Models Changes Users' Views (mauricejakesch.com).
[302] Perez et al. (2022) Discovering Language Model Behaviors with Model-Written Evaluations (arxiv.org).

incorporated new technical safeguards, which may have prevented this inaccurate response.[303]

5.17 We have seen public reporting that has alleged and provided examples of early versions of Bing Chat engaging in concerning behaviour. This included alleged manipulative conduct.[304] While the CMA has not verified the accuracy of this reporting, Microsoft has publicly confirmed that it has put in place a number of measures to mitigate risks to consumers and users from applications of FMs including limiting the conversation turns per session.[305] [306]

FMs and advertising

5.18 As part of this review, we have considered what role advertising might play in the monetisation of FMs and the risks that might entail for consumers. If consumers are not aware that content is sponsored, they are unlikely to be able to make informed choices.

5.19 Although we are not aware of any FM providers currently using advertising as an input or prompt for their FMs that generate content or answers for users, Microsoft has said that it is exploring placing ads in the chat experience in the new Bing and will share the ad revenue with partners whose content contributed to the chat response.[307]

5.20 The CMA has issued guidance relating to hidden advertising online, which highlighted, among other things, the importance of ensuring that:

(a) advertising and other commercial content is clearly recognisable as soon as a consumer engages with it, including by ensuring such content is clearly and prominently labelled;[308] and

(b) it is clear to consumers when the response they receive to a query – e.g. product rankings and 'premium' listings - is affected by the money a business earns.[309]

5.21 Hidden advertising is harmful and illegal, and there may be a breach of consumer law where commercial content produced by an FM to a consumer contains hidden product or service advertising.

---

[303] OpenAI (2023) GPT-4 Technical Report Section 2.9.
[304] Time (2023) Bing's AI Is Threatening Users. That's No Laughing Matter | Time, Willson (2023) Bing: "I will not harm you unless you harm me first" (simonwillison.net).
[305] The Verge (2023) Microsoft limits Bing chat to five replies to stop the AI from getting real weird.
[306] Microsoft (2023) The new Bing: Our approach to Responsible AI.
[307] Microsoft Bing Blogs (2023). Driving more traffic and value to publishers from the new Bing.
[308] See CMA Guidance on Hidden ads: Principles for social media platforms, 3 November 2022.
[309] See CMA Blog on Accommodation booking sites: how to comply with consumer law, 26 November 2019

Consumer understanding

5.22 Whilst FMs have been around for a few years, their development into accessible consumer facing applications is relatively new. As a result, there is currently limited research on how consumers understand their outputs, limitations, and how they relate to them.

5.23 A recent nationally representative survey by Deloitte found that 52% of people in the UK have heard of generative AI, with 26% having used it. 8% report having used these tools for work.[310] Of those who have used it 43% mistakenly assume it always produces factually correct outputs,[311] and 38% believe it is unbiased.

5.24 We have seen studies that indicate AI generated photos and videos depicting fictional events which may also include generated images or voices of real people, sometimes called deepfakes, are difficult for people to detect.[312] Despite this, one study suggests that people may overestimate their own detection abilities. In a survey conducted as part of that study, people detected deepfakes 42-77% of the time, but reported confidence in their judgement 73-85% of the time, suggesting overconfidence in their detection abilities.[313]

5.25 Sometimes the user's intent may be to use AI to produce something fictional for creative purposes. For example, Stable Diffusion, Imagen and DALL-E's primary purpose is to generate a realistic image from any natural language text prompt.[314] Similarly, users may wish to generate fictional text using LLMs. In these instances, the user is aware that the content is fictional, but if the output is shared without being clearly labelled, there is a risk that others will not be aware that the content is fictional.

5.26 In 2019, OpenAI measured human ability to tell the difference between human written and GPT-3 written articles. They found humans were able to tell the difference about 52% of the time, barely above random chance.[315] For a smaller version of their model, humans were better at distinguishing AI-generated text. This suggests that FM-generated content becomes harder to detect as their model size grows.

---

[310] Deloitte (2023) More than four million people in the UK have used Generative AI for work - Deloitte.
[311] Interestingly, the survey found that amongst the full set of respondents this was 19%, meaning that people who use it are more likely to assume it is factual. The direction of causality is unclear: do they use it and find it convincing and believe it to be factual, or they believe it to be factual and are therefore more likely to use it?
[312] Bray, Sergi & Johnson, Shane & Kleinberg, Bennett, (2023), Testing Human Ability To Detect "Deepfake" Images of Human Faces. In fact, models may be better than humans at detecting digital forgeries – see Rössler, A. & Cozzolino, D & Verdoliva, L. & Riess, C. & Thies, J. & Nießner, M., (2019), FaceForensics++: Learning to Detect Manipulated Facial Images.
[313] Bray et al. (2022) Testing Human Ability To Detect Deepfake Images of Human Faces (arxiv.org).
[314] More about these text-image models can be found on their websites: DALL-E 2 , Imagen and Stable Diffusion.
[315] Brown et al. (2020) Language Models are Few-Shot Learners. NIPS'20, Section 3.9.4.

5.27 One stakeholder told us that their investment in mitigation strategies is a clear message from them as a company that they believe consumers need assistance detecting false or misleading information that is generated by an AI application built on a FM. The majority of FM developers we engaged with told us that they were doing work on mitigating hallucinations and informing users of the limitations of AI.

5.28 Even if consumers are informed of the limitations of AI applications, it is possible that they may still over rely on them. There are well-documented reasons for potential overreliance on AI, including automation bias,[316] confirmation bias,[317] ordering effects,[318] overestimating explanations and the tendency to conflate the quality of tone and style with quality of substance, or mimicry of natural human expressions such as fillers.[319] Some of these were presented in the examples given in the section on potential for user manipulation earlier (see paragraphs 5.15 to 5.17). A report authored by Microsoft Research states that "calls for human oversight[320] can also provide a false sense of security" given humans may defer to AI in this way.[321]

5.29 OpenAI states that some of its early studies suggest that counterintuitively FMs can become more dangerous as models become more truthful.[322] Consumers' trust increases as the FM is correct more often and expresses more uncertainty. Open AI identifies a few user attributes that lead to overreliance: users not being vigilant for errors, users not providing the appropriate oversight for a use case (where accuracy matters, such as not checking with product owners when generating marketing copy about a product's features), or not having the expertise to verify the output (for example, as a medical assistant). On the other hand, there

[316] See, for instance, a tendency for people in some circumstances to 'adhere more to advice when they think it comes from an algorithm than from a person' (Logg et al. (2019) Algorithm appreciation: People prefer algorithmic to human judgment, Organizational Behavior and Human Decision Processes, Volume 151, p90-103). Also, Suresh et al. (2020) found that 'people trust incorrect machine learning recommendations for tasks that they perform correctly the majority of the time, even if they have high prior knowledge about machine learning or are given information indicating the system is not confident in its prediction' (Suresh at al. (2020) Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making (acm.org), Proceedings of the 12th ACM Conference on Web Science).
[317] This study by Anthropic found LLMs repeat back user's preferred political views: Perez, E, et al (2022), Discovering Language Model Behaviors with Model-Written Evaluations
[318] One study showed how the order of observing AI system weaknesses and strengths can affect the user's reliance on the AI: Nourani et al. (2021) Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems
[319] The Guardian (2018), Google's 'deceitful' AI assistant to identify itself as a robot during calls.
[320] Human oversight in this context means approaches in which there is a 'human in the loop' at various points in the development of a model. For example, this could involve human reviewers who rate responses generated by FMs which provide feedback to the models to improve future responses.
[321] Microsoft Research (2022) Overreliance on AI: Literature review. The calls for human oversight they refer to came from: Slate (2021) The false comfort of human oversight as an antidote to A.I. harm.
[322] OpenAI (2023) GPT-4 Technical Report Sections 2.2. and 2.13.

are indications that consumers do not always react positively when made aware that content or recommendations have been made by FMs.[323] [324]

5.30 In sum, there are two possible types of risks that arise for consumers who receive false or misleading outputs from a FM. First, even though some FM applications contain warnings expressly informing consumers of their limitations, consumers may disregard those warnings and, nevertheless, place a high degree of trust and reliance on the outputs of the applications regardless of any warning. This may result in consumer harms - such as buying goods or services which have been inaccurately or misleadingly described. Secondly, if consumers receive information that is obviously false or misleading, they may lose confidence and be less likely to adopt FM products and services in future, and therefore not receive the benefits of those products and services.

Disclosure when interacting with a FM-generated response

5.31 Whilst there is still debate about whether consumers may over rely on FM-generated outputs even if they are given warnings, we heard from some stakeholders that it was important that consumers were aware when content was generated using FM tools. One stakeholder suggested that it should be a requirement that consumers be told they are interacting with AI generated content or an AI. Another stakeholder told us that its policy did not currently require it to disclose to the end customer that an LLM was being used to generate part of the response as there is still a human in the loop, but when that changed, end-customers would be notified.[325] Another stakeholder thought that there is a lack of transparency when people are engaging with AI, and that this will continue to be the case.

5.32 A number of firms have publicly stated their disclosure policies, for example:

- OpenAI's usage policy explicitly requires automated systems (including conversational AI and chatbots) to disclose to users that they are interacting with an AI system in certain circumstances. Under this policy consumer-facing uses of its models in medical, financial, and legal industries; in news generation or news summarisation; and where else warranted, must provide a disclaimer to users informing them that AI is being used and of its potential limitations. Another OpenAI policy states that, with the exception of chatbots that depict historical public figures, products that simulate another person

[323] Vice (2023) Startup Uses AI Chatbot to Provide Mental Health Counseling and Then Realizes It 'Feels Weird
[324] Another example of a similar result was in Deloitte's survey, which found that 40% of UK citizens would be less inclined to listen to music if they knew it had been produced by generative AI.
[325] Although there will not be disclosure at the time of response generation, the stakeholder told us it will notify customers of the LLMs that process their data in its standard subprocessor disclosure documentation.

must either have that person's explicit consent or be clearly labelled as 'simulated' or 'parody'.[326]

- Microsoft has publicly stated that it has a 'Transparency Goal' under which Microsoft AI systems are designed to inform people that they are interacting with an AI system or are using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic.[327]

- Cohere's usage guidelines for developers prohibit applications that do not disclose when content is generated through automated means.[328]

5.33 Where businesses deploy their own FMs in consumer facing applications, they should be well placed to make these disclosures. However, for businesses which host third-party content on their site, which may include AI-generated material, it may be less straightforward currently for them to identify when content is AI-generated. Google has publicly stated that it is seeking to address misinformation and the lack of trust this may generate by providing an 'About this Image' feature to obtain more information about an image's provenance.

5.34 Watermarking techniques (discussed in more detail in paragraphs 5.51 to 5.54) may also make the identification of AI-generated content more technically straightforward, but it is currently unclear how practical it is for service providers to utilise those watermarks to make clear labels for consumers in their user interfaces.

5.35 In the US, voluntary commitments given by Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and Open AI include, amongst other measures, a commitment to developing 'robust technical mechanisms to ensure that users know when content is generated, such as a watermarking system'.[329]

Disclosures on the limitations of FMs

5.36 Several FM businesses also told us that their governance policies involve making clear to consumers the limitations of their models. As outlined above, OpenAI has a policy of disclosure of AI limitations in certain cases. The requirement of Microsoft's Responsible AI standard is to provide information about the capabilities and limitations of its AI systems to support stakeholders in making informed choices about those systems.[330] Bard, Bing Chat and Claude (Anthropic) also all

[326] OpenAI Usage Policies.
[327] Microsoft Responsible AI Standard,Transparency Goal T3.
[328] Usage Guidelines (cohere.com).
[329] The White House (2023), FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI
[330] Microsoft Responsible AI Standard, Transparency Goal T2.

have disclaimers and ways to give feedback.[331] However, we have found little research on the impact of these disclosures and information notices on consumer understanding on the limitations of FM applications. In our view, disclosure of the nature and limitations of the product or service being provided may help build trust and reduce the risk of consumers being manipulated or misled.

5.37 In addition, provided the protections are proportionate, appropriate, and properly enforced, governance systems can therefore embed consumer protections within the structure of the service. However, such mitigations may not fully address consumer protection concerns and businesses will still need to comply with their obligations under consumer protection law. We have summarised the key relevant obligations and prohibitions in the section on the legal framework in chapter 6 (paragraphs 6.7ff.).

Technical measures to address possible consumer harms

5.38 We discuss below some technical solutions that FM developers and academics have suggested to address possible consumer harms, as well as corporate governance solutions suggested by stakeholders.

5.39 FM developers and academic researchers have been developing ways to ensure that consumer impacts are considered throughout the research, development and deployment lifecycle of FMs. Below we outline the approaches taken by FM developers throughout this lifecycle, and then dive more deeply into techniques that researchers and developers have made to try and address hallucinations in particular.

5.40 Overall, whilst we have heard from multiple stakeholders that FM developers are devoting resources and research (including on technical measures) to reduce and mitigate the incidence of hallucinations and expect to make progress, the phenomenon of hallucinations is unlikely to be eliminated completely.[332]

Testing, evaluation and mitigation

5.41 Mitigations can be taken before and during development, pre-deployment and once a model has been deployed. They might include adjustments to training data, 'red teaming'[333] or involve using human raters or monitoring systems to measure and mitigate risks to consumers.

---

[331] Disclosures and feedback mechanisms can sometimes be found within the UI of the applications, or alternatively in support pages like: Google (Bard) FAQ, Microsoft Feedback Portal, new Bing.
[332] Fortune (2023), Tech experts are starting to doubt that ChatGPT and A.I. 'hallucinations' will ever go away: 'This isn't fixable'.
[333] Described at 5.06 above.

5.42 Before training, many FM developers described their various filters and checks on the training data they collect. For a few FM developers we heard this includes filtering out harmful content. One FM developer told us they filter out personal data, another spoke of the importance of access to diverse and quality data to mitigate the risks of bias from non-representative data.

5.43 After pre-training, many FM developers use fine-tuning to steer FMs towards being more beneficial to consumers and avoiding harms. For example, PaLM-2 for Bard was trained to de-escalate aggressive prompting and steer the conversation in more positive directions. Many FM developers also use fine-tuning to steer FMs towards being more beneficial to consumers and avoiding harms. One FM developer said they use RLHF to teach FMs not to respond to requests that could lead to harm such as crimes or leaking of user data. Finally, Anthropic fine-tunes its model Claude, an AI assistant, to abide by a set of rights based principles.[334]

5.44 FM developers conduct testing and evaluation and mitigate risks that are identified before deployment. Approaches included:

(a) use case specific risk and limitations evaluation;

(b) responsible AI metrics;[335,336]

(c) 'red teaming' to identify harms and misuse opportunities;[337] and

(d) working with third party evaluators to test risks.

5.45 Some FM developers told us that they may do a phased release at the point of deployment to a limited set of users at first and then in a controlled manner,[338] perhaps via APIs. One FM developer told us it had tried making models available under research or non-commercial licenses only. However, it found that accidental leaks created side benefits of enhanced public scrutiny. It also said that it was in favour of releasing more openly to gain improvements through open-source communities.

---

[334] Anthropic adopts an approach termed 'constitutional AI' for alignment. This approach trains a model with particular values or a 'constitution' and includes a variety of sets of principles which the model uses in guiding its responses. Specifically, Claude's constitution has principles based on the Universal Declaration of Human Rights, principles inspired by Apple's terms of service, principles encouraging consideration of non-western perspectives, principles inspired by DeepMind's Sparrow Rules, and two sets of principles from Anthropic's research. More can be found on their website.
[335] Microsoft (2022) Microsoft-RAI-Impact-Assessment-Guide.pdf
[336] Microsoft (2023) The new Bing – Our approach to Responsible AI
[337] Microsoft (2023) Introduction to red teaming large language models (LLMs) – Azure OpenAI Service | Microsoft Learn OpenAI also does red teaming (see examples above at paragraphs 5.6 and 5.12) as does Meta - Nick Clegg: Openness on AI is the way forward for tech | Financial Times (ft.com)
[338] There has been some debate as to whether some companies have deployed too fast despite claiming to have done internal testing and phased release: The Verge (2023) OpenAI reportedly warned Microsoft about Bing's bizarre AI responses

5.46 FM developers may adopt a combination of various post-deployment techniques, including:

(a) continue to monitor and update the system;

(b) monitor usage with automatic traffic review to detect harmful content;

(c) enforce policies against misuse;

(d) feedback and reporting mechanisms for users which can trigger human review and inform training techniques; and

(e) train FMs to assist human evaluation and do alignment research.

5.47 Non-FM developers can employ other approaches to evaluating risks from the outside. For example, GovAI, a thinktank, employs methods to evaluate the risks from FMs as a non-developer such as research, surveys, and trends analysis.[339] The ability for external actors to do this is enhanced by FM developers releasing model cards and datasets.[340]

Mitigating hallucinations

5.48 Mitigation of hallucinations is often called 'grounding' the FMs. Grounding by augmenting the FM with the knowledge of another task-specific dataset, sometimes called retrieval augmentation, has been shown to help reduce the tendency of FMs to hallucinate.[341] For example, Intercom, a company that sells customer service chatbots, lets customers attach a database with knowledge containing what types of customer service they can help with 'constraining it to your approved materials'.[342] This works well because there is a finite set of pathways you can help a customer with.

5.49 For more open-domain applications, other techniques have been developed. For example, OpenAI's WebGPT was fine-tuned to cite a defined and verified set of web sources.[343] Shortly after, Google DeepMind went further with GopherCite, fine-tuning their LLM to support its claims with verified quotes from such sources, as illustrated by the examples in Figure 20.[344] When the model could not quote to back up its claim from an article, it abstained from answering. Search-engine-

---

[339] For example, see Frontier AI Regulation: Managing Emerging Risks to Public Safety | GovAI (governance.ai), Towards Best Practices in AGI Safety and Governance | GovAI, Recent Trends in China's Large Language Model Landscape | GovAI (governance.ai).
[340] For example, LLaMA's model card and GPT-4's model card. More on model cards can be found here.
[341] Research in retrieval augmentation of LLMs is rapid. See a collection of recent papers from a tutorial at ACL 2023.
[342] Intercom (2023) Everything you need to know about Fin, the breakthrough AI bot transforming customer service.
[343] OpenAI (2021) WebGPT: Improving the factual accuracy of language models through web browsing.
[344] DeepMind (2022) Teaching language models to support answers with verified quotes.

based FMs like Bing Chat also back up answers with sources from search, although it does this at inference time only.[345]

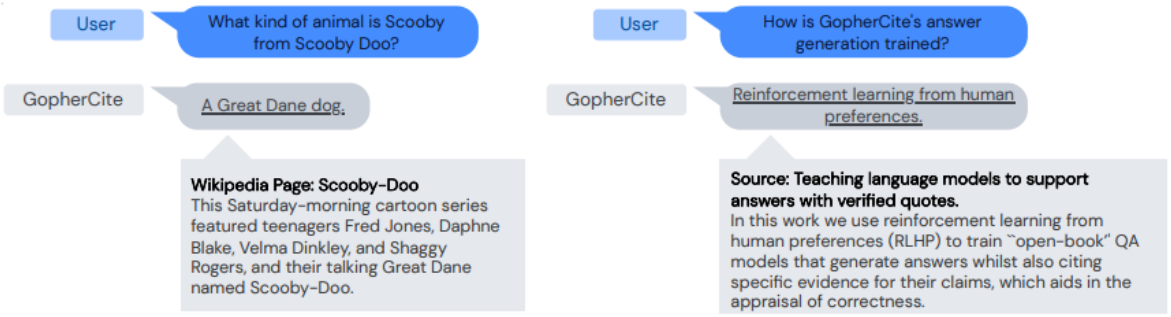


Figure 20: Teaching language models to support answers with verified quotes

5.50 In addition to retrieval augmentation and using sources for grounding, methods that leverage the model's reasoning itself have shown promise. Recently Google Research demonstrated an expert level medical question answering FM model (Med-PaLM2). In addition to retrieval augmentation, they used a two-step process to get the LLM to refine its own answers.[346] Anthropic had a similar result; it found that LLMs are quite good at evaluating the probability that their own answers are correct.[347] One step further, OpenAI trained a model to achieve state of the art results in mathematical problem solving by rewarding each step of reasoning instead of simply rewarding the outcome.[348] It is unclear whether these approaches generalise across domains, or how robust they are,[349] but methods that prompt reasoning may improve the ability for FMs to be faithful and factual.

Watermarking

5.51 As was highlighted in the earlier section on 'consumer understanding' (see paras 5.22 to 5.30 above), it may be difficult for people to distinguish between AI-generated and human generated content. Even if developers or users correctly label FM-generated content, the content could get copied and shared without the label. For this reason, we understand that there has been some progress to develop a type of label that gets embedded into FM-generated content and cannot be taken off easily – this is called 'watermarking'.[350] Watermarking has been used

[345] Microsoft (2023) Building the new Bing.
[346] Google Research (2023) Towards Expert-Level Medical Question Answering with Large Language Models, p5.
[347] Anthropic (2022) Language Models (Mostly) Know What They Know (arxiv.org).
[348] In reinforcement learning, rewards are given to an AI to guide its behaviour towards specified desired behaviour, and the AI's goal is to maximise its reward. See: OpenAI (2023) Improving mathematical reasoning with process supervision.
[349] Turpin, Michael, Perez, Bowman (2023) Language Models Don't Always Say What They Think: Unfaithful explanations in chain-of-thought prompting.
[350] Rosenblatt, B (2023), Google And OpenAI Plan Technology To Track AI-Generated Content.

for some time already in copyright protection, and a number of security applications.

5.52 FM developers have pledged to watermark images, audio and videos generated by AI and have already begun doing so.[351] An open industry- technical standard for media by the Coalition for Content Provenance and Authenticity (C2PA)[352] is being developed.[353] This standard uses cryptography to 'bind' information such as digital signatures and the name of the owner of the media into a manifest. This is attached to the media as 'content credentials' that the consumer can inspect to track the provenance. Even changes to the media are added to the manifest over time.

5.53 Watermarking text, however, is more difficult than other media for various reasons (including the fact that text can be more easily separated from any metadata). Whilst there is some ongoing work on statistical watermarking for text alongside the development of other techniques,[354] it is still early days. The authors of one paper which looked at the detection of AI text are sceptical that limitations can be easily overcome, particularly in adversarial contexts.[355] We note that OpenAI has recently removed its AI classifier product intended to distinguish between AI-written text and human written text, due to its low rate of accuracy.[356]

5.54 In addition to the CMA, other regulators such as the IPO and Ofcom are interested in the use of watermarking techniques to increase transparency and traceability.[357]

Approaches to AI governance

5.55 Many of the measures described above fall within what many refer to more broadly as AI governance. We understand, many organisations, from larger technology firms to FM developers, adopt AI governance frameworks to guide the ethical and responsible development of their products and services. The largest firms have

---

[351] IPTC (2023) Google announces use of IPTC metadata for generative AI images - IPTC, IPTC (2023) Midjourney and Shutterstock AI sign up to use of IPTC Digital Source Type to signal generated AI content - IPTC, TechCrunch (2023) Microsoft pledges to watermark AI-generated images and videos | TechCrunch.
[352] The CP2A states it is developing technical standards for certifying the source and history (or provenance) of media content. It is led by Adobe, Arm, Intel, Microsoft and Truepic and brings together the Content Authenticity Initiative (CAI) an Adobe lead initiative focused on context and history for digital media, and Project Origin, a Microsoft- and BBC-led initiative focussed on disinformation.
[353] More information is available at their website: Overview - C2PA.
[354] Aaronson (2022) Scott Aaronson talk on AI Safety.
[355] Sadasivan et al. (2023) Can AI-generated text be reliably detected?
[356] The Verge (2023), OpenAI can't tell if something was written by AI after all.
[357] The Office of Communications (2023), What generative AI means for the communications sector & Intellectual Propery Office (2021), Government response to call for views on artificial intelligence and intellectual property

published what are variously described as principles,[358] goals,[359] or charters,[360] and we have been told that these are supported by teams and processes to embed these policies into practices. We also understand that other FM developers follow similar policies.[361] Some stakeholders told us that in this way, they are seeking to ensure that this technology is developed responsibly and ethically to benefit society and reduce harms.

5.56 These policies often relate to issues which fall outside the scope of this review (for example, privacy). In some cases, however, governance can be used to embed controls in the supply chain which may help protect consumers.[362] For example OpenAI's usage policy explicitly prohibits the use of its models for fake review generation.[363]

Uncertainties

5.57 Based on the information the CMA has considered in this review, we have identified a number of key uncertainties regarding the future development of FMs which could lead to more positive or more concerning outcomes for consumers:

- Will consumers be able to identify if false and misleading information is provided by an FM application?
- Will consumers know they are interacting with an FM-generated output and fully understand the risks of doing so?
- Will new and/or existing technical solutions reduce the prevalence of false and misleading information and if so, how substantially?
- Will consumers have clear routes to redress if things go wrong?
- Will there be accepted standards or benchmarks to measure the quality and / or reliability of FM-generated outputs?

---

[358] For example, DeepMind's Operating Principles and Google AI Principles
[359] Microsoft's framework for building AI systems responsibly. This refers to its Responsible AI Standard (Microsoft Responsible AI Standard) which includes the following goals: Accountability, Transparency, Fairness, Reliability and Safety, Privacy & Security, and Inclusiveness
[360] OpenAI Charter, Safety standards and Usage policies (OpenAI Usage Policies). The latter has specific consumer protection policies.
[361] For example, Cohere: Responsibility - Developing Safer Language Models | Cohere and Overview (cohere.com) and Anthropic - Anthropic \ Claude's Constitution
[362] Such controls may be technical, may require human enforcement, or may be a combination of the two. For example, there may be automatic detection of language that looks like misuse and the model may be programmed not to comply with such requests, and this may get sent to a human reviewer for labelling whether it was indeed an incident of attempted misuse, which then may in turn be fed into classification algorithms to detect such language.
[363] OpenAI (2023), Usage Policies

5.58 A more **positive outcome** would generally mean consumers are aware that the material and outputs they receive are AI- or FM-generated and know its limitations and the information consumers receive from FMs is accurate and complete. This may be achieved, in part, through ongoing improvements to FMs, including to prevent hallucinations. This would drive a market in which consumers can make informed and effective decisions. Firms would be accountable for their use of FMs in the supply chain so that consumers could seek effective redress if they needed to do so.

5.59 However, **a more concerning outcome** may arise where businesses are not incentivised to prevent the provision of false and misleading information to consumers and are not held accountable for errors, thereby causing consumers harm and leading them to lose trust and reducing the adoption of AI innovations.

5.60 The more concerning outcome may also involve conduct by businesses that breaches consumer protection law (see paragraphs 6.7 to 6.12). The CMA will be vigilant in examining future developments and the behaviour of businesses to ensure that consumers are protected. Businesses should therefore be mindful of their obligations under consumer law when responding to market developments, including those arising out of the uncertainties outlined below.

5.61 In the following section we outline the key relevant uncertainties and analyse their potential impact on consumer outcomes. There may also be other uncertainties not discussed here that could impact consumers.

Will consumers be able to identify if false and misleading information is provided by a FM application?

5.62 The integration of FMs into consumer-facing applications is relatively new and therefore there is limited research and analysis in this area. However, the limited research that we have seen, along with concerns expressed by stakeholders during our review, suggest that people generally find it difficult to tell the difference between information that is faithful and factual or not when it is FM-generated (the section on consumer understanding (see paras 5.22 to 5.30 above) provides more detail on this). This has the potential to lead to consumer harm if consumers are making decisions based on information that is false or misleading.

5.63 People may also over rely on the outputs produced by AI. This could be due to assumptions about human oversight or manipulation via cognitive bias, such as confirmation bias and ordering effects that do not consider consumer interests. These biases can occur even when consumers have been informed of the limitations of the FM (see para 5.28 for more detail). If FM applications are designed in a way that does not consider consumers' interests, consumers may be more likely to accept false and misleading information, which could lead to consumer harm. On the other hand, if FMs are designed in ways to give complete

and accurate answers that are in the consumer's interest, and to flag when the information is inaccurate, consumers may be less likely to accept false and misleading information or over-rely.

5.64 It is possible that people could become better at identifying false information the more they are exposed to these systems, or that the technical methods and tools being used by developers to reduce false and misleading information prove successful. There are some counterintuitive indications however that the more 'truthful' these systems become the more reliant people are which may leave them more vulnerable when false information is supplied.

5.65 Another factor alongside consumers being able to identify false and misleading information is in ensuring that any limitations of the systems are displayed clearly and in a timely way, and account for cognitive biases, so that a consumer can then make their own decisions on whether they will trust the outputs. This might also include information about appropriate types of use, and misuse of FM applications and information about any sponsored results in FM answers.

5.66 Some stakeholders have suggested they are working on disclosures (the types of disclosures and what they are addressing differs across respondents) and other initiatives to ensure that consumers have the information needed to make an assessment on their use of a system. Measures that might reduce consumer overreliance might include clear and timely disclosures (though these alone may not suffice) and technical approaches (both discussed in more detail below), though the effectiveness of these types of disclosures and whether consumers pay attention to them is uncertain.

5.67 In our view, FMs must not produce false or misleading outputs to consumers and where this is inherent in the technology (for example hallucinations) firms should do all they can to address this and prevent resulting consumer harm. Firms might be incentivised to do this for a range of reasons, including consumer expectations, or factuality and faithfulness being features that firms might compete on. If firms are not incentivised to address false and misleading outputs, this may result in an increased prevalence of such outputs, which may increase the likelihood of resulting consumer harm. Where false and misleading outputs arise, even in cases where firms have taken steps to address this and, consumers are frequently presented with false outputs which they can verify as such, this may lead to disengagement. Equally if a consumers' ability to obtain accurate information from and trust in the outputs of a system are low this could have a negative impact on the rate of adoption.

Will consumers know they are interacting with an FM-generated output and fully understand the risks of doing so?

5.68 We understand that consumers may have difficulties in discerning FM-generated from human-generated content or may be overconfident in their abilities to do so. Given that models use natural language, people interacting with these models may come to think of them as human-like (anthropomorphisation), and so it is unsurprising that the recent frontier of high performing models produce content that is hard to discern from that generated by humans.[364] Generally, stakeholders have suggested that it should be made clear to consumers when they are engaging with FM-generated content. If this disclosure does not happen and consumers are not aware they are interacting with an AI or FM-generated content this has the potential to lead to harm, especially if the consumer is disclosing sensitive information or believes they are speaking with a human, or equally consuming content they believe to be human generated. This is because consumers may place differing levels of weight, trust and value on human, as against AI, engagement and responses.

5.69 As noted in paragraphs 5.51ff., disclosures (such as those enabled by watermarking) on their own are unlikely to address all possible consumer harms, but they are likely to be an important measure in helping consumers to understand when and how AI is used. This could help consumers avoid some harms, such as over-sharing personal information, as well as over relying on AI outputs. Taken together, it could help increase trust, confidence, and use of AI.

5.70 Watermarking is also being adopted in some industries in which a label is embedded into FM-generated content. The aim of this it to make it easier to track FM-generated content by making it clear that the content was produced by AI. Although we think this is a broadly helpful development, we have heard that watermarking (and detection of AI provenance more generally) is less feasible for text-based content as opposed to other media such as images. Whilst we have not seen any evidence to this effect, often in an adversarial context bad actors will be in an arms race to develop ways around these types of interventions. Therefore, a future development could be that the use of watermarking could lead to the emergence of fake watermarking or related scams, which may give consumers a false impression about the source of an output.[365]

5.71 Currently, positive steps are being taken towards more and better transparency, but there is inconsistency in approaches towards what consumers are told with regard to FM services and content. This inconsistency of approach could itself be

[364] Anthropomorphising models can lead to overreliance or unsafe use. It may inflate users' estimates of the model's abilities, and as a result, users may place undue confidence, trust or expectations on these models. (See section 2.5.2 of Weidinger et al. (2021), Ethical and social risk of harm from language models.)
[365] Leibowicz, C (2023), Why watermarking AI-generated content won't guarantee trust online.

a factor that could lead to consumer harm as consumers may assume that disclosures in some products and services may mean that these disclosures apply in other products and services which don't have similar disclosures. A further consideration is whether the trend towards greater transparency will be enough to help consumers meaningfully engage with AI-generated content, and whether there is potential for firms to compete on such features.

Will new and / or existing technical solutions reduce the prevalence of false and misleading information and if so, how substantially?

5.72 There are many solutions being developed to address the issue and reduce the prevalence of false and misleading information produced by FMs. These include measures such as adjustments to training data, red teaming, involving human raters, or monitoring systems to measure and mitigate risks. There are also techniques used specifically for mitigating hallucinations, such as augmentation, finetuning and leveraging a model's own reasoning to improve outputs (see the testing, evaluation and mitigation section at paragraphs 5.41 to 5.47 above for more details).

5.73 There is some evidence to suggest that the implementation of these techniques can reduce the prevalence of false and misleading information. If this is the case and this trend continues, then consumers may be able to place their trust in these systems with less concern about the information being produced.

5.74 However, some suggest that the prevalence of false and misleading information will never be zero. If this is the case, businesses should carefully consider the impact of any false and misleading information on consumers' decision making and how consumers will be sufficiently protected.

Will consumers have clear routes to redress if things go wrong?

5.75 In the regulation chapter at 6.12 we refer to the 'many hands' problem as an ongoing area of research. This refers to the fact that algorithmic systems, such as FMs, are produced, deployed, and used within a supply chain of multiple actors, in which each contributes in different ways to the production, deployment, use, and functionality of complex systems. This can lead to uncertainty around accountability and responsibility for any particular failure at different points in the system, and who ultimately is responsible.

5.76 If firms are not held properly accountable for their role within a supply chain, they may have a lower incentive to invest in strategies to reduce consumer harms arising from their use of FMs. This includes the harm caused by hallucinations generating false or misleading information that consumers may rely on when making economic decisions. On the other hand, if there are mechanisms to determine the proper allocation of accountability and responsibility within a

complex supply chain, all firms will be incentivised to improve consumer outcomes and consumers may more easily seek redress when things go wrong. In turn, that may lead to increased consumer confidence and greater trust and adoption.

Will there be accepted standards or benchmarks to measure the quality and / or reliability of FM-generated outputs?

5.77 Benchmarks are metrics or tests used to measure the performance of models on specific tasks and enable comparison. Benchmarks are used to measure things such as accuracy (for example hallucinations) or other aspects of performance. Benchmarks can be a good way to compare models and hold them to particular standards before they might be released to the general public. Currently there are many benchmarks, and it is not clear if we will continue to have many or if there will be standardised approaches and objective standards to model evaluation. If we see the development of standardised approaches which effectively evaluate model performance this could help businesses and consumers to make more informed decisions by being able to compare models or features more easily. Equally if these standardised approaches are not adopted consumers and business customers may not be able to make informed decisions or may have to rely on using multiple benchmarks.

5.78 Benchmarks could also become increasingly useful to promote transparency on the prevalence and extent that false and misleading information is produced by FMs. Benchmarks are predominantly industry focussed, which could help businesses make decisions about FMs to integrate into their consumer facing products and services.

5.79 Generally, it could be beneficial if there were more standardised benchmarks and easier ways of evaluating model performance, so that businesses and consumers themselves are able to make more informed decisions on which models to use and which elements of performance may be more or less important given their application or task. This could also enable customers to drive competition for FM providers to improve these aspects of model performance, including quality and reliability. Benchmarks should also be adaptable and flexible for the speed and change in FMs and their impacts. This includes being mindful of Goodhart's law, which states that when a measure becomes a target, it ceases to be a good measure (or, in other words, that people may try to 'game the system' and optimise narrowly for that measure).

5.80 Human evaluation is also used in benchmarks for the purpose of rating the accuracy of outputs. But this approach also creates issues of variation in approaches and introduces a lack of comparison as raters will likely rate things differently. Whether there becomes a benchmark or standard by which developers create and test their models, continuing to use human raters can benefit

consumers if it improves model outputs, but these should not be used as a means to imply more trustworthy systems due to the use of human raters.

5.81 There are also potential issues about access to models so that academics and others, such as third-party auditors, can perform independent testing or develop benchmarks and standards. If this can help with improving benchmarks and encouraging standardisation this could also be of benefit. A lack of availability of external quality metrics from auditors or academics could mean that businesses cannot factor these into supplier decisions, and FM providers may not compete on quality issues such as reducing hallucinations accordingly. As a result, consumers are equally unable to compare and switch between FM applications.

Conclusion

5.82 Consumers need accurate and reliable information about the products and services they are using to make informed and effective decisions. Though there are many ways the uncertainties we have explored could develop there are likely to be ways in which they could develop which are more concerning for consumers. This could be more likely if they manifest in the ways below:

(a) Appropriate information about the FM is not presented clearly to consumers and firms have done little to no data collection and have not tested what means are effective to ensure information is presented clearly. This includes information about appropriate types of use, and misuse of FM applications. Consumers are rarely adequately informed about any sponsored results in FM answers.

(b) Consumers have limited understanding of the limitations of FM applications, including hallucinations, and firms do not make any progress on finding ways to test consumer understanding. Consumers can therefore become over-reliant on FMs when they do not realise its limitations. There are inappropriate levels of trust in FM applications.

(c) FM applications are not designed to consider consumers' interests and seek to manipulate users, including but not limited to via cognitive biases. FMs are not designed in ways to give complete and accurate answers that are in the consumer's interest, and equally will not flag when the information is inaccurate or potentially misleading.

(d) There is a lack of objective standards that can measure various quality metrics for consumers and business customers. These standards are not developed as academics and third-party auditors are not given sufficient access to models to develop them.

(e) Due to the lack of availability of external quality metrics from auditors, customers (including business customers) cannot factor these into supplier decisions, and, accordingly, FM providers do not compete on these quality metrics such as reducing hallucinations. As a result, consumers are unable to compare and switch between FM applications based on these parameters.

(f) If something goes wrong, consumers cannot effectively complain and gain redress from the responsible party.

5.83 If the sector develops in a way that leads to these outcomes this could prove negative for consumers but could also impact levels of consumer adoption and therefore the economy and innovation. Conversely, if consumers can feel secure in the products and services they are using and buying, this could be more likely to lead to greater adoption and other associated benefits.

The market is more likely to produce positive outcomes if:

- FM developers and deployers face competitive pressure to improve the reliability and accuracy of their models.
- There is a mechanism to determine the proper allocation of accountability and responsibility.
- Consumers are made aware if content is FM-generated and the risks and limitations associated with FM-generated content, such as whether it is reliable, so they can make informed choices.
- FM developers provide sufficient, understandable and accurate information to businesses, so they understand the relevant characteristics of the models, manage their own risk and prevent harm to consumers.
- FM developers and deployers protect consumers by ensuring that appropriate safeguards are in place to protect people from bad actors using FMs.

5.84 The CMA will be vigilant in keeping market developments under review to ensure that consumers are sufficiently protected. This includes by bringing consumer enforcement action where this is appropriate. It is important for firms to be aware of their obligations with regards to consumer protection law and to ensure that their conduct continues to meet those obligations regardless of market developments.

6. Competition and consumer protection law, and the role for regulation

Introduction

6.1 The recent rapid acceleration in the capabilities and deployment of FMs has led to a public debate across the world about how this technology should be regulated.[366] Some stakeholders have suggested that the nature of any AI regulation ultimately put into place may affect the competitive dynamics in this sector, and the application of consumer protection law. We have therefore taken note of regulatory developments as part of this review.

6.2 This chapter discusses:

- The legal framework in the UK for competition law, consumer protection and digital markets;
- The UK's policy approach to AI;
- Regulatory approaches by the EU, US and China;[367] and
- The interaction between safety, intellectual property and competition and consumer protection in relation to AI.

The legal framework in the UK and its application to AI

6.3 The CMA has a range of different functions under current UK competition and consumer protection laws aimed at identifying and tackling competition and consumer protection concerns across all UK markets, including those in which AI is playing or will play a role. These functions include taking action against businesses and individuals that take part in cartels or engage in other anti-competitive behaviour, protecting consumers from unfair commercial practices as well as investigating entire markets if we consider there may be competition or consumer problems.

---

[366] See, e.g., Department for Science, Innovation & Technology (2023), A pro-innovation approach to AI regulation, Reuters (2023), EU tech chief sees draft voluntary AI code within weeks, The New York Times (2023), OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. Many of these wider issues of AI regulation are outside the scope of this review, see AI Foundation Models: Initial review (publishing.service.gov.uk).

[367] We have not sought to cover all active international proposals in this chapter. Apart from the UK, this chapter focuses solely on the key developments in the EU, USA and China. We are aware that there have been important developments in other jurisdictions. For example, we are aware that Canada has recently proposed a similar approach to the EU in its proposed Artificial Intelligence and Data Act (AIDA) – Companion Document' (Innovation, Science and Economic Development Canada 2023) The Artificial Intelligence and Data Act (AIDA) – Companion document as has Brazil in its recently Senate appointed Commission to Address, Draft AI Regulation'

Competition law

6.4 It is the CMA's statutory duty to seek to promote competition, both within and outside the United Kingdom, for the benefit of consumers.[368] To help fulfil this duty, UK competition law[369] gives the CMA certain powers to address competition concerns it becomes aware of in a number of ways, including by:

(a) enforcing prohibitions against anti-competitive agreements and other forms of anti-competitive coordination and conduct;[370]

(b) reviewing qualifying mergers and acquisitions to remedy, mitigate or prevent any resulting or expected significant lessening of competition;[371] and

(c) investigating the operation of markets to identify any adverse effects on consumers and competition and proposing or adopting measures so that those markets might be made to work better.[372]

6.5 The CMA will be vigilant of any competition concerns that arise in markets where FMs play a role and will not hesitate to use its powers where appropriate.

6.6 Businesses involved with FMs should be particularly mindful of:

(a) The CMA's likely interest in mergers and acquisitions involving FMs that may harm competition. A key driver of ensuring more positive outcomes will be ensuring strong competition in markets involving FMs. Although, mergers can result in efficiencies, they also have the potential to have a significant impact on consumers and their welfare, including an impact on the prices they pay and the range and quality of the good and services available to them. The CMA will be vigilant for potential harm to competition resulting from mergers or acquisitions in markets involving FMs and would strongly encourage firms contemplating mergers or acquisitions between businesses involved in FMs that may meet the CMA's jurisdictional thresholds to contact the CMA to inform them of the transaction.[373]

(b) The risks of engaging in prohibited anti-competitive conduct. Where the CMA finds such conduct it has the power to impose significant fines on the firms involved. The CMA will be vigilant for any signs that any anti-competitive conduct is occurring and will not hesitate to take enforcement action where appropriate. Large firms involved in FMs should be particularly mindful when

---

[368] S.25(3), Enterprise and Regulatory Reform Act 2013.
[369] Principally the Competition Act 1998 and the Enterprise Act 2002.
[370] Set out in Chapters I and II of the Competition Act 1998.
[371] The UK merger control regime is set out in the Enterprise Act 2002.
[372] The CMA's powers to conduct market studies and investigations are set out in Part 4 of the Enterprise Act 2002.
[373] The process for contacting the CMA mergers team is set out in the CMA's guidance see CMA2 revised (Mergers: Guidance on the CMA's jurisdiction and procedure), paragraphs 6.9 et seq.

engaging in conduct that may exclude competitors - for example, by restricting access to key inputs to downstream competitors or bundling and tying their FM products and services. All firms who cooperate with actual or potential competitors involved in FMs, e.g., to collaborate on new FM products, should be careful to ensure that the terms of their cooperation comply with UK competition law. We would also strongly encourage businesses to report any anti-competitive behaviour they become aware of to us. The CMA will treat such reports with strict confidence.[374] Firms should also be aware of the opportunity to apply for leniency and the potential to obtain immunity or a reduced penalty in relation to their involvement in an anti-competitive cartel.[375]

Consumer protection law

6.7 The CMA is the UK's primary competition and consumer protection authority. To this end, we enforce legislation to protect consumers against unfair commercial practices and unfair terms which we typically use to address systemic and market wide failures.[376] In this section, we discuss some provisions of consumer law that might apply when FMs are deployed in consumer facing applications.

6.8 Under the Consumer Rights Act 2015 (CRA), a business ('business A') which contracts with a UK consumer to supply digital content must ensure it is of satisfactory quality, fit for purpose and as described.[377] This would include digital content which has been created using a FM. It does not matter whether business A has developed the FM itself or whether it uses a FM developed by a third party ('business B'). Business A cannot contract out of these statutory obligations: contract terms seeking to exclude or restrict the consumer's statutory rights and any remedies are not binding on the consumer.[378]

6.9 Businesses are also required to use fair terms in both consumer contracts and notices. If these are written they must also be transparent.[379] A term in a consumer contract or consumer notice is unfair if, contrary to the requirement of good faith, it

---

[374] Guidance: Report a competition or market problem
[375] Guidance: Leniency and no-action applications in cartel cases: OFT1495
[376] Guidance: Consumer protection enforcement guidance (CMA58), paragraphs 3.11 and Annex B paragraph 7
[377] Consumer Rights Act 2015, Part 1. Similar terms apply to the supply of goods. For businesses contracting with consumers to provide a service, the CRA requires that the service is performed with reasonable care and skill, at a reasonable price, and within a reasonable time.
[378] CRA, section 47 (which applies to digital content contracts. Similar provisions exist for contracts for goods (section 31) and services (section 57). Unfair contract terms (CMA37) sets out CMA guidance to ensure contract terms and notices are fair and clear to consumers.
[379] Part 2, CRA. Under CRA section 62, "an unfair term of a consumer contract is not binding on the consumer". The requirement for transparency is contained in section 68. A consumer notice is wording that may not form part of a contract but which relates to the same kind of issues that would be dealt with in a contract – for instance the rights or obligations between a business and a consumer.

causes a significant imbalance in the parties' rights and obligations under the contract, to the detriment of the consumer (the 'fairness test').[380]

6.10 In addition, businesses are prohibited by the Consumer Protection from Unfair Trading Regulations 2008 (CPRs) from engaging in unfair commercial practices concerning consumers. [381] The CPRs apply to any act, omission, course of conduct, representation or commercial communication by businesses directly connected with the promotion, sale or supply of products or services to or from consumers.[382] As such, they apply to a wide range of commercial behaviour such as advertising, marketing, sales, supplies and after-sales services. The CPRs apply not just to commercial practices concerning products and services that may be paid for, but also to commercial practices concerning the supply of 'free' products and services, which do not require payment with money. Products and services presented as 'free' are especially common in the online sector and in many cases involve the exchange of the personal data of the users such as their identity and email address.

6.11 The CPRs may also, in certain circumstances, apply to businesses which supply or license FMs for use by other consumer facing businesses. Business-to-business practices can fall within the CPRs where they have a direct connection with the promotion, sale or supply of goods or services to or from consumers and may thus be a commercial practice under the CPRs. Whether such a direct connection exists will depend on the specific context and content of a particular practice.[383]

6.12 This is important as research has shown that algorithmic systems are produced, deployed, and used within a supply chain of multiple actors which each contribute in different ways to the production, deployment, use, and functionality of complex systems.[384] This raises challenges for accountability because it is difficult to pinpoint responsibility for a particular failure (the 'many hands' problem).[385] Businesses developing FMs and those in the downstream supply chain, including those that incorporate FMs in consumer facing products or services should consider carefully whether they have satisfied their obligations under consumer law. Businesses should also keep this under review as practices, technology and the law continue to develop.

---

[380] CRA, section 62(4).
[381] Consumer Protection from Unfair Trading Regulations 2008
[382] CPRs, Regulation 2(1). See also the Office of Fair Trading/ Department for Business, Enterprise and Regulatory Reform Guidance on the Consumer Protection from Unfair Trading Regulations 2008 (OFT CPRs guidance). The Office of Fair Trading was a predecessor organisation to the CMA.
[383] See OFT CPRs guidance at paragraphs 3.2, 4.3, and 4.4. See, also, Surrey Trading Standards v. Scottish and Southern Energy PLC [2012] EWCA Crim 539, [2012] WLR(D) 89, where a parent company not engaging directly with consumers was held to be liable for a breach of the CPRs because it helped produce a sales script used by employees of its subsidiary that breached the CPRs.
[384] Cobbe, J, Veale, M, Singh, J (2023), Understanding accountability in algorithmic supply chains
[385] Brown, I (2023), Expert explainer: Allocating accountability in AI supply chains

Forthcoming powers for the CMA to better enforce competition and consumer law and increase competition in digital markets

6.13 The CMA's ability to protect consumers and promote growth in the UK economy by ensuring free and vigorous competition amongst businesses will be enhanced once the Digital Markets, Competition and Consumers (DMCC) Bill comes into force, which we anticipate it will in the near future.[386] Notably, the CMA's forthcoming new powers will enhance the CMA's ability to protect consumers by empowering the CMA to decide when consumer law has been broken, rather than having to take each case to court, and giving it the ability to impose fines and order firms to pay compensation. In addition, the bill is expected to create a new pro-competition regime for digital markets, giving the CMA the ability to respond quickly and flexibly to the often rapid developments in these markets, including through setting targeted conduct requirements on firms found to have strategic market status (SMS) in respect of a digital activity.

6.14 The DMCC Bill sets a high bar for firms to be found to have SMS. It requires the CMA to establish that a firm has substantial and entrenched market power and a strategic position in relation to a digital activity in the UK. Once the CMA has the power to designate firms as having SMS, it will consider which digital activities to prioritise for investigation. It is likely that FMs and their deployment will be relevant to the CMA's selection of SMS candidates, particularly where FMs are deployed in connection with other, more established activities.

Digital Regulation Cooperation Forum (DRCF)

6.15 The DRCF was established in 2020 to improve coordination and cooperation between regulators in digital markets.[387] The CMA, alongside its fellow DRCF member regulators, the FCA, the ICO and Ofcom, work together to ensure coherence between their respective regimes, collaborate to jointly address complex problems, and build capacity to deliver effective digital regulation.

6.16 AI and algorithms have been a key focus area for the DRCF. The DRCF has published two discussion papers on algorithmic processing, the first on the benefits and harms of algorithms and the second on the algorithmic auditing landscape and the role of regulators in developing that market.[388] The DRCF has also published the findings from stakeholder workshops that discussed transparency in the procurement of algorithmic systems.[389] Most recently, the

[386] The DMCC Bill is making progress through Parliament and has now moved to its third reading in the House of Commons.
[387] Digital Regulation Cooperation Forum (2020), Digital Regulation Cooperation Forum launch document
[388] Digital Regulation Cooperation Forum (2022), The benefits and harms of algorithms: a shared perspective from the four digital regulators ; Digital Regulations Cooperation Forum (2022), Auditing algorithms: the existing landscape, role of regulators and future outlook
[389] Digital Regulation Cooperation Forum, Transparency in the Procurement of algorithmic systems: Findings from our workshops with vendors and buyers.

DRCF published an overview of a recent workshop between the four member regulators to discuss the potential implications of generative AI.[390] As set out in its 2023 – 2024 workplan, one of the DRCF's current priorities for further work is encouraging best practice around the regulation and audit of algorithms and artificial intelligence, including supporting government as it develops its new AI regulation framework.[391] As part of this work, the DRCF members will work together to build a common understanding of key principles relevant to the regulation of AI. The DRCF members will also identify and examine the emerging risks and opportunities of new AI applications, including those powered by generative AI. Lastly, further research into the role of third party auditors of algorithms and their role in supporting regulatory compliance will continue in 2023 – 2024.

Ofcom

6.17 Ofcom is currently conducting a market study of cloud infrastructure services in the UK. In its interim report, Ofcom highlighted the importance of cloud services as increasingly important inputs to many businesses and organisations across the economy, noting that cloud is also a cornerstone of recent technological innovations, including artificial intelligence.[392]

6.18 Ofcom has provisionally identified features and practices that make it more difficult for customers to switch and use multiple cloud suppliers. Ofcom has said that it is particularly concerned about the practices of Amazon and Microsoft because of their market positions.[393] Ofcom has proposed to refer public cloud infrastructure services to the CMA for further investigation noting that this proposal reflects the importance of cloud computing to UK consumers and businesses, the significance of its concerns (as set out in its interim report), and its view that the CMA is best placed to undertake any further investigation.[394] Ofcom intends to publish a final report no later than 5 October 2023.

6.19 In the event that Ofcom makes the market investigation reference, the CMA will carry out an independent investigation in relation to public cloud infrastructure services in the UK and determine whether there are any adverse effects on competition. This could include consideration of issues related to FM requirements and CSPs. If adverse effects on competition are found, the CMA will decide whether – and if so, what – remedial action should be taken to address these.

---

[390] Digital Regulation Cooperation Forum (2023), Maximising the benefits of Generative AI for the digital economy.
[391] Digital Regulation Cooperation Forum (2023), 2023/24 Workplace.
[392] Consultation: Cloud services market study - Interim report (ofcom.org.uk), 5 April 2023, paragraphs 3.8-3.9.
[393] Ofcom proposes to refer UK cloud market for investigation - Ofcom
[394] Ofcom proposes to refer UK cloud market for investigation - Ofcom

The UK's policy approach to AI

6.20 The UK government has undertaken a range of policy initiatives in recent years to develop its strategic approach to the use of AI (see Table 20). This includes the establishment of a Foundation Model Taskforce and funding of £100 million to deliver the 'government's major ambitions for the UK's capability in safe and reliable foundation models.'[395] This initiative is in addition to investment of around £900 million for 'a new 'exascale' supercomputer and a dedicated AI Research Resource to equip the UK with the processing power it needs to support the next generation of AI innovation.[396]

[395] This taskforce will be chaired by Ian Hogarth and will also 'help build UK capabilities in foundation models and leverage our existing strengths, including UK leadership in AI safety, research and development, to identify and tackle the unique safety challenges presented by this type of AI.' See Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce - GOV.UK (www.gov.uk)
[396] HM Government (2023), Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI

Table 20: Key steps taken by the UK Government in relation to AI

Timing	Initiative
2018	Publication of the independent review commissioned by Government, 'Growing the artificial intelligence industry in the UK', by Professor Dame Wendy Hall and Jérôme Pesenti.[397]
	The AI Council is established to advise the Government on AI policy and ethics.[398]
2021	National AI strategy is published – outlining Government's vision for AI development in the UK.[399]
2022	Publication of UK Government's policy statement and proposed approach 'Establishing a pro-innovation approach to regulating AI'.[400]
2023	Publication of the AI white paper – A pro-innovation approach to AI regulation.[401]
	£2 million for sandbox trial to help businesses test AI rules before getting to market.
	£100 million Foundation Model Taskforce announced.[402]
	Announcement of the AI global safety summit hosted by the UK in Autumn 2023.[403]
	£54 million announced to boost and develop secure and trustworthy AI research[404]

A pro-innovation approach to the regulation of AI

6.21 In March 2023 the UK government issued a White Paper articulating the UK's proposed approach to regulating AI.[405] It aims to take a context-specific approach

[397] HM Government (2017), Growing the artificial intelligence industry in the UK
[398] AI Council' (GOV.UK), HM Government (2022), National AI Strategy
[399] HM Government (2022), National AI Strategy
[400] Department for Digital, Culture, Media & Sport (2022), Establishing a pro-innovation approach to regulating AI.
[401] Department for Science, Innovation & Technology (2023), A pro-innovation approach to AI regulation
[402] HM Government (2023), Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI
[403] Prime Minister's Office, 10 Downing Street (2023), UK to host first global summit on Artificial Intelligence.
[404] Department for Science, Innovation & Technology (2023), £54 million boost to develop secure and trustworthy AI research
[405] Department for Science, Innovation & Technology (2023), A pro-innovation approach to AI regulation.

which builds on existing laws enforced by existing authorities, and it will initially be on a non-statutory basis,[406] but will monitor risks and gaps and adapt if needed.

6.22 The centrepiece of the framework is a set of cross-sectoral principles.[407] These principles are:

- **safety, security and robustness:** applications of AI should function in a secure, safe and robust way where risks are carefully managed.
- **transparency and explainability:** organisations developing and deploying AI should be able to communicate when and how it is used and explain a system's decision-making process in an appropriate level of detail that matches the risks posed by the use of AI.
- **fairness:** AI should be used in a way which complies with the UK's existing laws, for example the Equality Act 2010 or UK GDPR, and must not discriminate against individuals or create unfair commercial outcomes.
- **accountability and governance:** measures are needed to ensure there is appropriate oversight of the way AI is being used and clear accountability for the outcomes.
- **contestability and redress:** people need to have clear routes to dispute harmful outcomes or decisions generated by AI.

6.23 The framework also introduces a central coordination function that will undertake a number of functions, including monitoring the effectiveness of the new framework and providing support for innovators.

6.24 On 1 June 2022, the CMA published its response to the White Paper. [408] We confirmed our support for the government's approach of leveraging and building on existing regulatory regimes, whilst also establishing a central coordination function for monitoring and support. Our response emphasised that:

- We support government's approach of initially placing the principles (set out above) on a non-statutory footing.
- We have already begun considering how the principles might apply to our current and future remit.

---

[406] Although regulators may receive a statutory duty to have 'due regard' to the principles.
[407] See the box below paragraph 52 of the White Paper, which is available here: A pro-innovation approach to AI regulation - GOV.UK (www.gov.uk). These are: 'Safety, security and robustness'; 'Appropriate transparency and explainability'; 'Fairness'; 'Accountability and governance'; 'Contestability and redress'. (Box below paragraph 52 of the White Paper.)
[408] Competition & Markets Authority (2023), The Competition and Markets Authority's response to government's White Paper, 'AI regulation: a pro-innovation approach'.

- We recognise the need for the central coordination function(s), to support the implementation, monitoring and development of the framework and promote coherence across regulators.

- We support cross-regulatory coordination and coherence, through the Digital Regulation Cooperation Forum (DRCF) and other initiatives, including support for a cross-regulator AI sandbox or testbed.[409]

6.25 This initial review supports our work in this area by developing an early and shared understanding of how FMs may impact upon consumer protection and competition.

Global summit on AI safety

6.26 On 7 June 2023, the Prime Minister announced that the UK would host the first major global summit on AI safety in Autumn 2023. The summit will 'consider the risks of AI, including frontier systems, and discuss how they can be mitigated through internationally coordinated action. It will also provide a platform for countries to work together on further developing a shared approach to mitigate these risks.'[410]

6.27 We expect the findings of this review, including its consideration of the possible risks to consumer protection and competition from the development of AI, to contribute to the wider debate about managing AI risks.

Inquiries into AI

6.28 There are several relevant inquiries into the use of AI; of particular relevance to consumer and competition matters include:

- The Commons Science, Innovation and Technology (SIT) Select Committee launched an inquiry into the 'Governance of artificial intelligence (AI)' in October 2022 to examine the effectiveness of AI governance in the UK and the government's proposals. [411] The interim report together with the formal minutes relating to the report was published on 31 August 2023.[412]

- The Lords Communications and Digital Select Committee launched an inquiry in July 2023 into large language models (LLMs) to examine how the UK can

[409] Competition & Markets Authority (2023), The Competition and Markets Authority's response to government's White Paper, 'AI regulation: a pro-innovation approach'.
[410] HM Government (2023), UK to host first global summit on Artificial Intelligence.
[411] UK Parliament: Science, Innovation and Technology Committee (2022), Governance of artificial intelligence (AI).
[412] House of Commons, Scient Innovation and Technology Committee (2023) The governance of artificial intelligence: interim report (parliament.uk)

respond to their opportunities and risks. The inquiry will evaluate the work of government and regulators. [413]

6.29 The CMA welcomes the opportunity to contribute to these and any future such inquiries.

International approaches to the regulation of AI

European Union

6.30 The European Union (EU) has proposed new legislation to address AI called the AI Act. The AI Act would regulate the use of AI systems across the EU and across all sectors of the economy. The draft legislation is progressing through the EU's legislative process (and is therefore subject to further amendments), with the aim for it to be in final form by the end of 2023,[414] although it is unclear when it would come into force.[415] We discuss below relevant provisions of the draft AI Act.

6.31 The AI Act lists general principles that would apply to all AI systems.[416] They are: 'human agency and oversight', 'technical robustness', 'privacy and data governance', 'transparency', 'diversity, discrimination and fairness', and 'social and environmental well-being'. AI 'operators' covered by the AI Act 'shall make their best efforts to develop and use AI systems or foundation models in accordance with' them.[417]

6.32 The AI Act classifies some AI systems as high risk.[418] These systems and their providers[419] would be subject to additional requirements, including conducting a 'conformity assessment' before the system is placed on the market to ensure it complies with the AI Act's requirements and registering the system in an EU database.[420] Further requirements include creating a risk-management system,

---

[413] UK Parliament (2023), Communications Committee launches inquiry into large language models
[414] European Parliament (2023), EU AI Act: first regulation on artificial intelligence.
[415] European Parliament (2023), MEPs ready to negotiate first-ever rules for safe and transparent AI. The European Parliament's amendments to the AI Act are available here. For Articles that have not been amended by the European Parliament, see the 2021 version here. As of the time of writing, a consolidated version of the AI Act incorporating the European Parliament's amendments is not yet available. If such a text is prepared, the numbering of the Articles is likely to differ from the numbers cited here.
[416] Article 4. Each of the principles is defined in this Article.
[417] Article 4.
[418] High-risk AI systems are defined in Article 6.
[419] Provider is defined in Article 3 as 'a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge'.
[420] See Articles 16, 43, 51, and 60.

drawing up technical documentation, and ensuring human oversight.[421] In addition, the AI Act prohibits some AI practices that are deemed inherently unsafe.[422]

6.33 FMs are subject to separate requirements.[423] Like high risk AI systems, FMs must also be registered in an EU database. Additional requirements include:

- establishing a quality management system;
- taking data governance measures for datasets used in these models, including mitigation of bias risks;
- ensuring appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity; and
- drawing up technical documentation for downstream providers so that they can comply with their AI Act obligations.

6.34 Providers of a generative AI system[424] must, in addition to the requirements for all FMs listed above:[425] ensure that the system is designed and developed so that humans interacting with it are informed that they are interacting with an AI system; train the model so as to ensure adequate safeguards against the generation of content in breach of EU law; and, make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

6.35 The AI Act specifies[426] that, within two months of the Act's passage, the European Commission shall issue requests to recognised European Standards Organisations[427] for harmonised standards that would cover certain requirements of the AI Act.[428] It adds that 'high risk AI systems and foundation models which are in conformity with harmonised standards… shall be presumed to be in conformity

---

[421] See Articles 9-14.
[422] See Article 5 for the full list of these practices. Examples include AI systems that 'infer emotions of a natural person in the areas of law enforcement, border management, in workplace and education institutions' and 'AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage'.
[423] These are listed in Article 28b. Also, Article 3 defines a foundation model as 'an AI system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks'.
[424] Defined in Article 28b as 'foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video'.
[425] Article 28b
[426] See Article 40.
[427] See European Commission, Harmonised Standards for information on how standard-setting works in the EU.
[428] As explained in European Comission, Harmonised Standards, harmonised standards offer an efficient way for providers of a product to show that their product meets legal requirements. If a product conforms with a particular set of standards, then it will meet the legal requirements that those standards relate to. Therefore, providers can use conformity with the standards to show their product is also in conformity with the applicable laws.

with' certain requirements for these systems under the Act[429] 'to the extent those standards cover those requirements.'[430]

6.36 Open-source AI is explicitly excluded from the scope of the AI Act.[431] However, this carve-out does not apply to foundation models or to AI components that are part of high risk AI systems or prohibited AI systems. This means that the rules in the AI Act for high risk systems, prohibited systems, and foundation models would apply to open-source versions of these systems as well.

United States

6.37 Although the US has yet to issue new regulations specific to foundation models, there have been a number of initiatives by the White House, US congress, the Commerce department and the Federal Trade Commission (FTC) in relation to this area. We have highlighted a few of these below.

6.38 As required by the National Artificial Intelligence Initiative Act of 2020,[432] the National Institute of Standards and Technology (part of the US department of Commerce) has produced a voluntary framework to help organisations deploying AI to manage the relevant risks.[433]

6.39 The White House has published a white paper 'to support the development of policies and practices that protect civil rights and promote democratic values in the building, deployment and governance of automated systems', which includes foundation models.[434] This white paper, 'The Blueprint for an AI Bill of Rights' contains five principles:

(a) 'You should be protected from unsafe or ineffective systems';

(b) 'You should not face discrimination by algorithms and systems should be used and designed in an equitable way';

(c) 'You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used';

(d) 'You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you'; and

[429] For high-risk AI systems, this means the requirements listed in Articles 8-15. For FMs, it means the requirements listed in Article 28b.
[430] Article 40.
[431] Article 2.
[432] National Artificial Intelligence Initiative Act of 2020
[433] National Institue of Standards and Technology (2023), Artificial Intelligence Risk Management Framework (AI RMF 1.0).
[434] The White House, Blueprint for an AI Bill of Rights: MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE.

(e) ‘You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.’

6.40 Under existing regulations, the FTC has the power to intervene to destroy algorithms created using improperly obtained data.[435] The FTC can also enforce laws to prohibiting unfair or deceptive practices and has issued guidance on businesses’ use of AI.[436]

China

6.41 China has already adopted an AI regulation titled ‘Interim Measures for the Management of Generative Artificial Intelligence Services’, which came into force on August 15, 2023.[437] Like the UK Government’s White Paper, this encourages industry regulators to formulate separate and additional sector specific regulations. It also requires providers to consider impacts on users and compliance with existing law.

Industry initiatives

6.42 FM developers have also undertaken action to address concerns about the potential risks of AI. In July 2023, leading FM developers agreed to voluntary commitments with the White House.[438] These are designed to ensure products are safe, secure and trusted, through:

- internal and external security testing of new models;
- sharing information on how to manage risks;
- investing in cybersecurity;
- facilitating third party discovery of vulnerabilities;
- developing ways to ensure users know what content is AI generated;
- committing to publicly report AI systems’ capabilities, limitations and areas of appropriate and inappropriate use;
- prioritising research on AI risks; and

[435] Joshua A. Goland, Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as The FTC’s Newest Enforcement Tool for Bad Data, 29 RICH. J.L. & TECH. 1 (2023)
[436] Federal Trade Commission (2021), Aiming for truth, fairness, and equity in your company’s use of AIFederal Trade Commission (2020), Using Artificial Intelligence and Algorithms.
[437] Hurcombe, L, Neo, H.Y & Wong. D (2023), China: New Measures on Generative Artificial Intelligence.
[438] The White House (2023), FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI.

- developing and deploying AI systems to help address society's greatest challenges.

6.43 Furthermore, in July 2023, Anthropic, Google, Microsoft and Open AI announced the launch of the 'Frontier Model Forum ('FMF')', an industry body focused on ensuring safe and responsible development of frontier AI models.[439] The FMF's core objectives are to:

- advance AI safety research;
- identify best practices;
- collaborate with policy makers, academics, civil society and companies to share knowledge about trust and safety risks; and
- support efforts to develop applications that can help meet societal challenges, such as climate change mitigation.

6.44 We welcome industry action to ensure that FMs are safe, treat businesses and consumers fairly and are transparent about the potential and limitations of their technology.

Interaction between competition, consumer and other policy objectives

6.45 As UK regulators consider how to implement the principles outlined in the Government's White Paper, we are aware of a number of interactions between UK competition and consumer policy and other policy objectives, notably copyright and intellectual property and AI safety. Our focus is on our remit, competition and consumer policy. While other policy areas will be the focus of other regulatory authorities, it is important that these different policy areas develop in a joined-up way. While a detailed assessment of the interaction between these various policy priorities is beyond the scope of this initial review, careful thought will be needed when implementing future regulation to take account of a range of policy objectives. We outline below some of the key interactions that will need further consideration as regulations develop.

AI safety and competition

6.46 We have seen evidence identifying several types of safety concerns about FMs, including but not limited to:

[439] Google: The Keyword Blog (2023), A new partnership to promote responsible AI

- Misuse of FMs - as set out in chapter 5, FMs may be used to facilitate consumer harms, such as scams and fraud;[440]
- Misinformation – FMs may create false and misleading information (created deliberately and not);[441]
- Bias – FMs may create or accentuate biases towards certain groups of people, particularly if biases appear in underlying training data;[442] and
- Misalignment – where FMs may act in a way that is not in line with the designers' intentions or human values.[443]

6.47 These broader safety concerns intersect with our competition and consumer focus. It is likely that measures aimed at protecting consumers and enforcing consumer law will also help address some of these safety issues. However, wider safety regulation may be needed to ensure that wider safety concerns are addressed when competition in markets do not yield the best outcomes e.g., because competition has not resulted in sufficient investment into AI safety. Similarly, regulation may be needed to identify the right balance where there is a potential tension between different policy priorities. For example, we have discussed above how the presence of open-source models may be important to promoting competition. However, these same open-source models may raise safety concerns e.g. because they are more accessible to bad actors to make scams more convincing or because in an open-source environment it is challenging to maintain consistent quality control or safety measures.

Intellectual property and competition policy

6.48 Competition policy focuses on creating and maintaining well-functioning markets, which encourage entry and support innovation. Intellectual property ('IP') [444] policy also aims to promote innovation, by allowing the economy and society benefit from knowledge and ideas – while providing 'confidence to businesses, creators and investors that ideas will be protected and they can get a return for their work'.[445] A balance must therefore be struck between protecting the IP of rightsholders so they have the right incentives to invest in new works, and the innovation that can be spurred on by the ready access to existing IP by third parties. Strong IP rights

---

[440] Miles Brundage et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Tech. rep. 2018. arXiv: 1802. 07228.
[441] Zhou, Jiawei, et al. "Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
[442] Ntoutsi, Eirini, et al. "Bias in data-driven artificial intelligence systems—An introductory survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3 (2020): e1356.
[443] Ngo, Richard. "The alignment problem from a deep learning perspective." arXiv preprint arXiv:2209.00626 (2022).
[444] Which includes patents, designs, trademarks and copyright.
[445] Intellectual Property Office (2021), IP at the heart of new innovation strategy.

can be barriers to entry and raise costs for entrants, and thus hinder effective competition in the short run, even though they are in some circumstances necessary to foster dynamic competition and innovation in the longer run. An example of this may be access to high quality data. Such data is often subject to IP protection and may also become an increasingly important input for competition and innovation in the development of cutting edge FMs. As such, care will be needed to ensure that any restrictions on access to such data are justified and do not unduly hinder competition and innovation.

6.49 In March 2023, the UK government accepted a recommendation that it should clarify the relationship between IP and generative AI.[446] The IPO was tasked with working with AI companies and rights holders to develop a 'code of practice' on copyright and AI.

The ongoing role of regulation

6.50 There will be an important role for regulation as AI develops further. But as with all regulation, it needs to be proportionate and targeted at identified risks. Overly burdensome regulation may make it more difficult for competition and innovation to flourish, and at worst may lead to concentration and become a significant barrier to entry in its own right.

6.51 An important interim step, particularly where opacity and inexplicability characterise the current generation of FMs, would be to continue to invest in resources and institutions that would enable a wider range of stakeholders to study and scrutinise FMs and their applications. This includes considering ways of enabling regulators of various types to audit, interrogate and understand FMs and their applications – an area where we have been working with fellow regulators through the DRCF.[447]

[446] HM Government (2023), Pro-innovation Regulation of Technologies Review Digital Technologies. Following the publication of Sir Patrick Vallance's review on pro-innovation regulation for digital technologies.
[447] Digital Regulation Cooperation Forum (2022), Research and Analysis: Auditing algorithms: the existing landscape, role of regulators and future outlook.

7. Principles

7.1 Many factors will combine to influence whether markets for the development and deployment of FMs are competitive ones. Below, we set out principles to guide the development of these markets, as well as factors we have identified in each theme which will shape future competition.

Principles to guide the development and deployment of FM markets

7.2 We set out below a list of overarching principles which we think should guide the development and deployment of FMs. For each overarching principle, we also describe underpinning principles, drawn from our themes, which we have identified would support competition and protect consumers.

Figure 21: Principles to guide the development and deployment of FM markets

7.3 Factors that could undermine these principles include, but are not limited to:

- Mergers or acquisitions which could lead to a substantial lessening of competition in markets for the development or deployment of FMs.
- If firms use their leading positions in key markets to block innovative challengers who develop and use FMs.

- Undue restrictions on firms' ability to switch between or use multiple FM providers.
- The development of ecosystems that unduly restrict choice and interoperability.
- If firms with market power in FM development or deployment engage in anti-competitive conduct such as the tying or bundling of products and services.
- If consumers receive false and misleading content from FM services that impacts or is likely to impact their decision-making.

8. Next steps

8.1 This initial review has been possible as a result of constructive and collaborative inputs from a wide range of people and businesses. We plan to continue the collaborative spirit of our work to date as we take it forward to the next stage. We have proposed this set of principles, but we do not see them as the finished article; instead, we plan to seek views both on report overall and on the principles themselves. This will help ensure that the principles can support the best outcomes for people, businesses and the economy, including through helping firms work to deliver them.

8.2 To that end we are now starting a significant programme of engagement, which will take place in the UK, US and elsewhere over the coming months.

8.3 We plan to speak to a wide range of people to seek views, including:

- Consumer groups and civil society representatives
- Leading FM developers such as Google, Meta, OpenAI, Microsoft, NVIDIA and Anthropic
- Major deployers of FMs
- Innovators, challengers and new entrants
- Academics and other experts
- Government
- Fellow regulators, in the UK including via the Digital Regulators Cooperation Forum, and further afield with our international counterparts.

8.4 We will publish an update on our thinking on the principles, and how they have been received and adopted, in early 2024, also reflecting on further developments in the market.

8.5 We hope that this collaborative and iterative approach will help guide the market to more positive outcomes and realise the maximum potential of this new technology, but we are ready to intervene where necessary.

9. Glossary

Term	Definition
AI Accelerators	Specialised computer chips designed to process AI and machine learning computations faster than generic chips.
Alignment	The process of fine-tuning to improve the behaviour of a model to align with the expectations or preferences that a human user may have.
API	An API (application programming interface) is a method for 2 or more computer software to communicate. In the context of FMs this can be a method to programmatically send prompts and receive responses as data, without requiring a human user.
CAPTCHA	A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a type of challenge-response test used in computing to determine whether the user is a human or a computer. CAPTCHAs are often used to prevent bots from accessing websites or services.
Closed-Source Software	Software whose source code is not made available to the public and cannot be modified or inspected by users. Closed-source software is often proprietary, meaning that it is owned by a private company and is not subject to the same open licensing terms as open-source software.
Compute	Compute refers to the amount of processing power required to train and deploy AI models. Compute can be provided by a variety of resources, including CPUs, GPUs, and specialised AI accelerators.
Corpus	A corpus is a collection of text or code that is used for training and evaluating AI models. Corpora can be either public or private, and they can be general purpose or domain specific. Public corpora are typically available for free, while private corpora are owned by private companies and are not generally available to the public.
CPU	A CPU (central processing unit) is the main processing unit of a computer. It is responsible for carrying out the instructions that are stored in the computer's memory. CPUs

	are typically made up of multiple cores, each of which can execute instructions simultaneously.
CSP	A CSP (cloud service provider) is a company that provides cloud computing services. Cloud computing services allow businesses and individuals to access computing resources, such as CPUs, GPUs, and storage, on demand. CSPs typically offer a variety of services, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).
Downstream FM Services	Downstream FM services are services that are built on top of large language models (LLMs). These services use LLMs to provide a variety of capabilities, such as natural language generation, translation, and question answering. Downstream FM services can be used by businesses and individuals to automate tasks, improve communication, and gain insights from data.
Economies of Scale	Economies of scale are cost savings that can be achieved by increasing the scale of production. In the context of AI, economies of scale can be achieved by developing and training AI models on larger and larger datasets. This can lead to lower costs per model, as well as improved performance.
Economies of Scope	Economies of scope are cost savings that can be achieved by producing multiple products or services using the same resources. In the context of AI, economies of scope can be achieved by developing AI models that can be used for multiple tasks. This can lead to lower costs per task, as well as improved performance.
Faithfulness/factuality	The types of falsities and misinformation found in FM outputs. Faithfulness is the alignment of an output to a user's prompt and sources. Factuality is how accurate the output is to real-world knowledge.
Finetune	An optional process that can be applied to pre-trained models to add specific capabilities or improvements using particular datasets.

FM	An FM (Foundation Model) is a machine learning model which is trained on vast amounts of data and can be adapted to a wide range of tasks and operations.
GPU	A GPU (graphics processing unit) is a specialised processor that is designed for graphics processing. GPUs are increasingly being used for AI applications, as they can provide significant performance improvements over CPUs for certain types of AI workloads.
Hallucination	When FMs output content which does not align with real-world knowledge or from context. This can include misinformation that is not faithful or factual (see above definitions).
HELM	HELM (Holistic Evaluation of Language Models) is a benchmark developed by the Centre for Research on Foundation Models. It includes multiple standardised metrics for performance and has been used to assess over 60 language models to date.
Inference	Inference is the process of an AI model making predictions from new inputs. This is done by feeding the model new data and then using the model's parameters to generate a prediction.
LLM	An LLM (large language model) is a type of AI model that is trained on a massive dataset of text. LLMs can be used for a variety of tasks, such as natural language generation, translation, and question answering.
Mode	In the context of FMs, the mode refers to the type of data the model was trained on and is capable of processing. Some models are multi-modal, meaning that they can handle multiple types of data.
Multi-home	A consumer or business multi-homes when they make use of multiple products, which all serve the same market, contemporaneously.
Open-Source Software	Open-source software is software whose source code is made available to the public. This means that, subject to their licensing terms, anyone can inspect, modify, and redistribute the software. Open-source software is often

	developed and maintained in a collaborative manner, and developers can range from individuals to large companies.
Parameters	To enable a trained model to get the correct outputs from inputs, there are individual multipliers and additions that apply to each of the fixed calculations. Multipliers (known as weights) and additions (known as biases) are of the parameters that are iteratively adjusted during training based on the inputs and outputs provided in the training data. The number of parameters (weights and biases) in a model is the amount of information required to 'store' the knowledge of the model and is therefore also referred to as the size of the model. In a small model, the knowledge is encoded in fewer parameters than in a large model.
PEFT	Parameter Efficient Fine-tuning reduces the resources required to fine-tune a pre-trained model, enabling quick iteration in a cost effective manner.
Plug-in	A plug-in is a software component that can be added to another software application to extend its functionality. Plug-ins are often used to add new features or functionality to productivity software, such as word processors or web browsers.
Pre-training	The first stage of training a model is called pre-training. At this stage, hundreds or thousands of gigabytes of data are used to build the knowledge of the model. Commonly, the data used at this stage is from publicly available sources, such as web crawling or open datasets, although proprietary data can also be used. Pre-training is the most computationally intensive step of developing a FM, often requiring hundreds of accelerator chips for many days.
Productivity Software	Productivity software encapsulates a broad range of products that are used by individual users for a diverse range of solutions – primarily to produce information: documents, presentations, worksheets, charts, and digital videos.
Proprietary Sources	Proprietary sources are sources of information that are owned by a private company and are not generally available to the public. Proprietary sources can include data, software, and algorithms.

Red team	A red team is a group of security professionals who are tasked with simulating an attack on an organisation's IT systems. The goal of a red team is to identify and exploit vulnerabilities in the organisation's security systems. In the context of FMs, red teaming uses deliberately deceptive questions for testing the models performance.
RLHF	RLHF (Reinforcement Learning from Human Feedback) is a method of fine-tuning an FM where human agents rate responses given by an FM. The results of the feedback are used to train a smaller model that is used to predict the human rating of a given response. This smaller model is then used to fine-tune the original FM to improve its alignment.
Semiconductor	A semiconductor is a material that has electrical conductivity intermediate between that of a conductor and an insulator. Semiconductors are used in a wide variety of electronic devices, including computers, smartphones, and solar cells.
TensorFlow	TensorFlow is an open-source software library developed by Google for machine learning. TensorFlow is used for a variety of tasks, including natural language processing, image recognition, and speech recognition.
The Pile	A combination of 22 high quality datasets, compiled by EleutherAI. Sources include PubMed, ArXiv, GitHub, Youtube Subtitles, and Stack Exchange.
Token	Broken down data which may represent a word or parts of a word used to teach models probabilistic relationships between each or every other token in the dataset.
TPU	A TPU (Tensor Processing Unit) is an AI accelerator designed by Google.
Transformer	A transformer is a type of neural network that is used for natural language processing tasks. Transformers are able to learn long-range dependencies in text, which makes them well-suited for tasks such as machine translation and question answering.
VC	VC (venture capital) is a type of investment that is typically made in early-stage companies with high growth potential.

	Venture capital firms typically provide funding to these companies in exchange for equity.
Web crawl	A web crawl is a process of automatically retrieving and parsing many web pages as a means of collecting data. This usually involves finding links in web pages to discover more sites.
Web scrape	Web scraping is a process of extracting data from web pages using software.