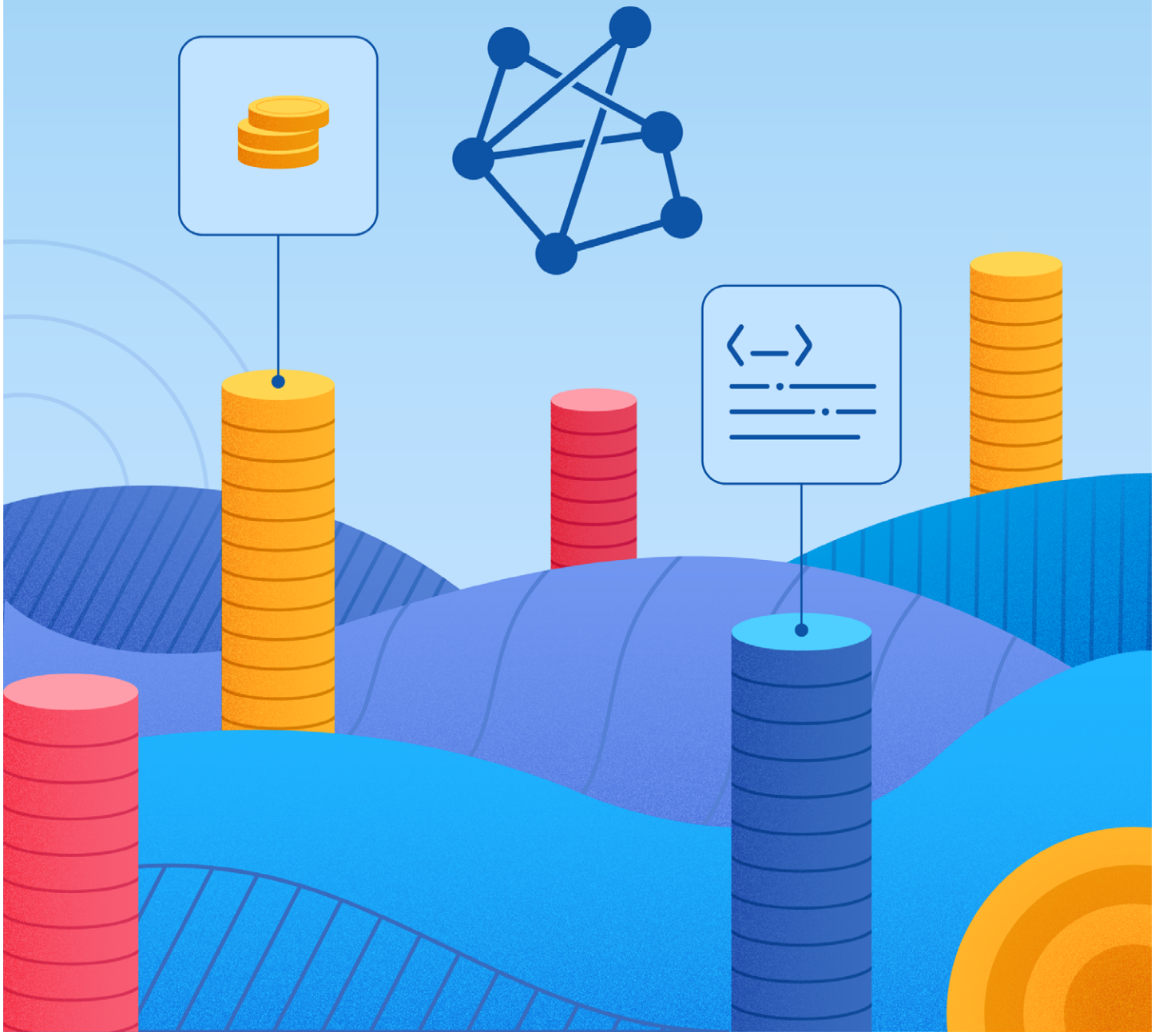
AI Value Creation and Open-Weight Models

Evidence from software production, consumer banking and private equity



BertelsmannStiftung

© Bertelsmann Stiftung, Gütersloh
November 2026

Publisher
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256, 33311 Gütersloh
www.bertelsmann-stiftung.de

Authors
Dr. Felix Sieker, Bertelsmann Stiftung
Max Schulze, Leitmotiv Digital

Responsible
Dr. Felix Sieker, Bertelsmann Stiftung

Editing
Barbara Serfozo, Berlin

Cover
© Paul Feldkamp, Bielefeld

Layout and Typesetting
Ines Meyer, Gütersloh

Rights
The **text** of this publication is licensed under the Creative Commons Attribution 4.0 International License. You can find the complete license text at:
https://creativecommons.org/licenses/by/4.0/legalcode.en



The **cover illustration,** the **graphics** and the **logos** are protected by copyright and are not covered by the above-mentioned CC license; they must not be used.

Recommended citation style
Schulze, Max & Sieker, Felix (2026). AI Value Creation and Open-Weight Models: Evidence from software production, consumer banking and private equity. Bertelsmann Stiftung. Gütersloh

DOI 10.11586/2026144

AI Value Creation and Open-Weight Models

Evidence from software production, consumer banking and private equity

Max Schulze & Felix Sieker

BertelsmannStiftung

Contents

Preface 6

Executive summary 7

1 | Introduction: Europe's missing demand-side AI strategy 8

2 | Public AI and open-weight models 9
2.1 | Open-weight models within the Public AI framework 9
2.2 | Competition and concentration 9
2.3 | Accountability, control and digital rights 11

3 | Competitive advantage from applying AI 12
3.1 | Value creation 12
3.1.1 | Cost advantage 13
3.1.2 | Differentiation 13
3.2 | Case selection 14
3.2.1 | Software production by freelance software engineers 15
3.2.2 | Consumer-facing banking 15
3.2.3 | Private equity 15
3.3 | Methodology 16
3.3.1 | Research design 16
3.3.2 | Semi-structured interviews 16
3.3.3 | Limitations 16

4 | Applying the framework to three AI use cases: empirical analysis 17
4.1 | Software production 17
4.1.1 | Value creation 17
4.1.2 | Capabilities 18
4.2 | Consumer-facing banking services 19
4.2.1 | Value creation 19
4.2.2 | Capabilities 22
4.3 | Private equity 25
4.3.1 | Value creation 25
4.3.2 | Capabilities 25
4.4 | Cross-case conclusion 27

5 | Prices, capability and barriers to open-weight adoption **28**
5.1 | The price of the models examined in Chapter 4 29
5.2 | The broader capability-price relationship 30
5.3 | Sources of sustained competitive advantage 31
5.4 | Why lower prices have not led to faster adoption 33

6 | Discussion **34**

7 | Risks to the current open-model price advantage **36**
7.1 | The advantage depends on licenses that permit competitive inference 36
7.2 | The advantage depends on an inference market with multiple providers 37
7.3 | Current prices may not be durable 38

8 | Conclusion and outlook **40**

Preface

Europe needs to give the adoption of artificial intelligence greater weight in its economic strategy. Much of the recent debate has centered on expanding computing capacity, developing advanced models and strengthening European AI providers. These investments can improve Europe's technological capabilities. Their wider economic contribution depends on whether firms can put these capabilities to productive use. Europe therefore also needs to understand where AI can create value for firms and what enables them to realize it.

The European Commission's Apply AI Strategy recognizes the importance of uptake across industries and the public sector. Broader adoption depends on the organizations that have to put AI to use. Firms need to identify activities in which AI can create value, assess the capabilities required and integrate suitable AI systems into their operations. A stronger demand-side perspective would help connect Europe's investments in AI to these requirements and direct support toward the obstacles firms encounter in putting the technology to use.

With this study, we seek to shed light on how firms adopt AI. We examine software production, consumer banking and private equity and distinguish between two ways in which AI can create economic value. It can contribute to a cost advantage by reducing the resources required to perform an activity, or support differentiation by improving products and services in ways that customers value. We then examine which models offer the capabilities needed for these applications.

The cases show that firms are using AI primarily to increase productivity. They also highlight the largely unrealized potential for differentiation. Developing better products and services could open up new sources of revenue and growth. This potential deserves a more prominent place in companies' AI strategies, particularly where efficiency improvements are becoming common across competitors.

Open-weight models broaden the options available to firms pursuing these objectives. The comparisons in this study indicate that leading open-weight models are already good enough for many relevant applications and are available at substantially lower prices. Where licenses permit, firms can also adapt them to their own requirements and choose between providers serving the same underlying model. Open-weight models therefore expand firms' strategic options, without making access to the most capable proprietary frontier model a condition for many valuable applications.

European policymakers should protect the market conditions that make these choices possible. Firms investing in AI need continued access to capable models on competitive terms. More restrictive licenses or a decline in the number of independent providers serving them could narrow their options and increase dependency. Preserving competition and the availability of capable open models should therefore be part of Europe's AI strategy.

We extend our sincere thanks to Max Schulze from Leitmotiv for his valuable insights and contributions to this work. We hope this study contributes to a more practical discussion about how Europe can turn its investments in AI into wider economic value and about the conditions firms need to make effective use of the technology.

Dr. Kirsten Witte
Director
Center for Sustainable Communities
Bertelsmann Stiftung

Executive summary

Europe's AI strategy has focused heavily on expanding the supply of computing capacity and advanced models. This report adds a demand-side perspective. It asks where firms can create economic value with AI, which model capabilities they actually require and whether those capabilities need to come from proprietary frontier models. To assess value creation, the report distinguishes between two mechanisms: (1) cost advantage, where AI reduces the resources required to perform an activity, and (2) differentiation, where AI improves products or services in ways that customers value. It applies this framework to software production, consumer banking and private equity and then examines which models can provide the capabilities required in these use cases.

Across the three cases, the evidence is considerably stronger for cost advantage than for differentiation. In software production and private equity, AI increases productivity by reducing the time required for coding, research, screening and other recurring activities. In consumer banking, the automation of routine interactions provides a plausible route to lower service costs. By contrast, there is little evidence so far that AI allows firms to charge higher prices or creates durable differentiation. Differentiation could ultimately become a more transformative source of value if AI enables firms to offer products and services that customers value more highly, but the cases provide little evidence that this potential has yet been realized.

Leading open-weight models provide sufficient capabilities for many of these valuable applications. On benchmarks covering software engineering, customer service in banking and financial analysis, the strongest open-weight models perform close to leading proprietary models. At the same time, the selected open-weight models are substantially cheaper to access. In software production and financial analysis, the leading open-weight alternatives cost less than one-tenth as much per output token as the strongest proprietary models. For many applications, firms therefore do not need the most capable proprietary models to obtain the capabilities required to create value.

For Europe, open-weight models are particularly important because they can help firms retain more of the value they create with AI. Where they provide the required capabilities at lower inference prices, European firms can spend less on model access and retain a larger share of the value created through higher productivity. Open weights can also make AI supply more competitive. Where licenses permit independent hosting, several providers can serve the same model, creating opportunities for European cloud and inference providers and application developers. This enables firms to gain greater choice among suppliers and more control over how models are adapted and deployed.

Realizing this advantage depends on the services and capabilities around the models. Proprietary providers remain attractive because they combine strong models with subscriptions, established tools, integrated systems and trusted infrastructure. Switching costs and limited organizational capabilities can make cheaper open-weight alternatives harder to adopt. Europe therefore also needs the infrastructure, integration services and deployment expertise that allow firms to use open-weight models effectively.

The current market configuration is highly favorable to Europe, but it may not last. Its advantages depend on capable models continuing to be released under licenses that permit independent commercial use, on several inference providers remaining able to compete and on sufficient funding for continued model development. Tighter licenses, greater concentration among inference providers or inadequate financing for future models could weaken the combination of low prices, provider choice and strong capabilities that exists today. A follow-up paper examines whether the current financing and revenue arrangements can sustain this market structure.

1 | Introduction: Europe's missing demand-side AI strategy

Europe is still seeking its own path to AI development and adoption. Current initiatives, including the AI Continent Action Plan, place substantial emphasis on expanding the supply side: increasing compute capacity, supporting frontier model development and strengthening European AI labs.[1] The underlying assumption is that greater access to infrastructure and advanced models will accelerate the development and diffusion of AI across the economy.

Expanding access to computing capacity and advanced models is an important part of strengthening Europe's AI capabilities. Yet it does not explain where AI can create value in the economy or which capabilities firms require to realize that value. It also directs attention toward frontier AI: very large, general-purpose models that are often proprietary and currently provided by a small number of companies through APIs, subscriptions and cloud platforms. Greater access to these systems may increase adoption, but it can also reinforce the concentration of technological and economic power.

This report adds a demand-side perspective by examining how AI could create value in specific sectors and use cases. It examines how firms apply AI and asks how doing so can affect their economic performance, for example, by reducing costs or improving products and services. It then identifies the model capabilities required to realize that value and assesses which models can provide them.

This perspective allows us to examine two central propositions. First, firms can create economic value with AI, either by reducing costs or by differentiating their products and services. Second, they do not necessarily need to depend on proprietary frontier models to do so. Where open-weight models provide the capabilities required for a particular use case at a lower price, firms can reduce their expenditure on model access and retain more of the value created.

The relevant comparison is therefore use case-specific. Some applications require the strongest available general-purpose model. Others can be served adequately by an open-weight model with narrower capabilities. Model choice should therefore follow from the capabilities and requirements of the particular use case. A demand-side analysis makes these differences visible and avoids treating model scale as a precondition for economic usefulness.

The report is structured as follows: Chapter 2 situates open-weight models within the wider Public AI framework. Chapter 3 develops the analytical framework, explains the case selection and describes the methodology. Chapter 4 applies the framework to three empirical settings – software production by freelance software engineers, consumer-facing banking and private equity – and assesses the capability gap between proprietary frontier and open-weight models. Chapter 5 compares the prices of frontier and open-weight models with comparable capabilities. Chapter 6 discusses the implications for a European AI strategy. Chapter 7 sets out the risks to the current price advantage of open-weight models and introduces the questions that a subsequent paper examines quantitatively.

1 Foundation models, a term coined by researchers at Stanford University, have been defined as "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." Bommasani, R. et al. (2022), *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, revised 12 July. https://crfm.stanford.edu/report.html. Frontier models are the most advanced foundation models.

2 | Public AI and open-weight models

This chapter builds on the Public AI framework developed in a white paper published by the Bertelsmann Stiftung.[2] Public AI refers to AI systems developed with transparent governance, public accountability, open and equitable access to core components and clearly defined public functions. The white paper treats compute, data and models as interdependent layers of the AI stack and identifies the development of open models and the wider ecosystem around them as one of three complementary pathways toward Public AI.

Open-weight models are important within this framework because they can change who is able to provide, inspect and adapt AI capabilities.[3] For Europe, two consequences are particularly relevant. First, open-weight models can make the provision of AI services more contestable and reduce dependence on a small number of vertically integrated suppliers. Second, they can widen the scope for independent scrutiny and give adopting organizations greater control over model adaptation and deployment. These effects concern both economic competition and public accountability, as outlined in this chapter.

2.1 | Open-weight models within the Public AI framework

In the Public AI White Paper, the proposed open-model pathway combines a highly capable "capstone model" with an ecosystem of smaller or domain-specific models derived from open foundations. The pathway also depends on access to the compute, data, open-source software and technical expertise needed to develop, adapt and deploy these models.[4]

Open-weight models include large general-purpose systems as well as smaller and more specialized derivatives. Their significance does not follow from their size. It follows from the fact that, where the license permits and the technical requirements can be met, the same model can be adapted and served through different infrastructure arrangements, including independent inference providers and self-hosted deployments. Depending on the use case and deployment model, they may offer lower operating costs, greater control over data and infrastructure and a stronger outside option in negotiations with suppliers.

2.2 | Competition and concentration

At present, enterprise adoption largely follows a frontier-provider logic. Powerful, broadly applicable systems are supplied through APIs, subscriptions and integrated cloud services. This route is attractive because it gives firms rapid access to advanced general-purpose capabilities without requiring them to build and operate the full technical stack themselves.

2 Sieker, F. et al. (2025), *Public AI – White Paper*, Gütersloh: Bertelsmann Stiftung. https://doi.org/10.11586/2025040.

3 The Public AI framework uses "open models" as an umbrella term for open-weight and open-source models. This report refers primarily to open-weight models, reflecting their greater prevalence. An open-weight release makes the model's trained parameters available under a license that determines whether and how the model may be used, modified, redistributed or commercially hosted. A fully open-source release goes further by also making the relevant code, training and evaluation data, and documentation available under free and open-source licenses.

4 Sieker, F. et al. (2025), *Public AI – White Paper*, Gütersloh: Bertelsmann Stiftung. https://doi.org/10.11586/2025040.

The same integration, however, can create economic dependency.[5] A single provider may control the model, the technical interface, the surrounding development tools and the infrastructure through which the system is accessed. Once a firm has built workflows around one ecosystem, switching may require changes to prompts, tool integrations, evaluation procedures, security controls and operating processes.[6] Coupling model access, tooling and infrastructure can therefore raise the cost of exit and strengthen the provider's position over prices, functionality and access conditions.

Open-weight models can alter this structure by separating model development from inference provision, since the same model can be deployed through different infrastructure arrangements. These can range from local or self-hosted infrastructure to independent inference providers such as "neoclouds." These providers can use open-weight models without paying license fees where the applicable license permits. Because multiple inference providers can run the same open-weight model and sell tokens, they can compete on price as well as service quality.

Marketplaces for model inference make this separation visible. On one such marketplace, OpenRouter, several providers may offer access to the same open-weight model at different prices and levels of performance. Because users can compare inference providers serving the same model, providers are able to compete directly on the price and quality of inference. Compared with proprietary frontier AI services, where users generally cannot freely choose among inference providers, this structure can create greater competition on price and service quality.[7]

Irrespective of the model being served, inference providers can differentiate themselves through the quality, reliability and trustworthiness of the infrastructure they provide. OpenRouter, for example, displays token generation speed, time to first token, measured availability and provider trust for each provider serving the same model.[8]

While frontier AI models from Anthropic and OpenAI have also become available through third-party inference providers, including Amazon Bedrock, Microsoft Azure Foundry and Google Vertex AI, these

5 See internal memo by OpenAI's chief revenue officer: "As the platform gets embedded, switching costs rise. As customers run more workflows through the system, OpenAI becomes harder to replace and more central to how work gets done. That is how we move from product vendor to operating infrastructure. … This is the flywheel we should be building around: better models drive more usage, more usage drives deeper integration, deeper integration drives multi-product adoption, and multi-product adoption makes us harder to replace." Quoted in Field, H. (2026), "Read OpenAI's latest internal memo about beating the competition — including Anthropic," The Verge, 13 April. https://www.theverge.com/ai-artificial-intelligence/911118/openai-memo-cro-ai-competition-anthropic.

6 "The Commission is also alert to the risk that customers' choice of AI technology or AI solutions may be restricted to offerings of their cloud provider or its partners and result in lock-in, given that switching cloud service providers seems to remain challenging as noted above (e.g., because of limitations in interoperability and data portability and various software licensing practices)." European Union (2025), *Competition in the Provision of Cloud Computing Services*, note to the OECD, DAF/COMP/WD(2025)39, 19 June, p. 7. https://one.oecd.org/document/DAF/COMP/WD(2025)39/en/pdf.

7 Open-weight models can be hosted and served by any inference provider, which then competes on price and speed. OpenRouter, for example, lists NVIDIA's Nemotron served by multiple competing providers ranked by price and throughput. See OpenRouter (n.d.), "Nemotron 3 Super – API Pricing & Benchmarks," provider comparison. https://openrouter.ai/nvidia/nemotron-3-super-120b-a12b#providers. Accessed 7 October 2026. This is possible because such models are released under permissive open-weight licenses: Nemotron is published under the NVIDIA Open Model License (Nemotron 3 additionally under the Linux Foundation's OpenMDW-1.1), which permits commercial use, fine-tuning, self-hosting and redistribution, including offering the model as a hosted service or API, without license fees or royalties, and does not claim ownership of model outputs. See NVIDIA (2025), "NVIDIA Nemotron Open Model License," last modified 15 December. https://www.nvidia.com/en-us/agreements/enterprise-software/nvidia-nemotron-open-model-license/. Other widely served open-weight/open-source models (e.g., Mistral NeMo, Z-AI GLM) are available on comparably permissive terms.

8 OpenRouter's per-provider profiles surface, and rank on, latency (time to first token), throughput (tokens/second), measured uptime/availability and provider trust, making inference-serving quality directly comparable across providers of the same model. See, for example, OpenRouter (n.d.), "Nemotron 3 Super," provider comparison.

arrangements do not necessarily allow inference providers to compete on price. For example, Anthropic applies the same prices across third-party inference providers through its partnership program and so-called Claude Consumption Units.[9]

Independent inference providers can compete to serve the same model without first financing frontier model development, while adopting firms gain an outside option to vertically integrated suppliers. Where an open-weight model is sufficiently capable for a given task, this competition can lower user prices and reduce dependence on any single provider.

2.3 | Accountability, control and digital rights

Open-weight models can also strengthen independent scrutiny. Access to model weights allows regulators, researchers and adopting firms to run a defined model themselves, reproduce behavioral evaluations and test it under conditions chosen independently of the developer. Fully open-source models go even further. They allow full access to training data, code and documentation, permitting closer examination of data provenance, development choices and systematic bias.

They also allow organizations to adapt a model. Where the license permits, firms and public institutions can fine-tune or distill it for a particular language, sector or workflow. For firms, this can improve performance on a specific use case, allow smaller and cheaper models to perform tasks that would otherwise require a larger general-purpose model and combine the model with proprietary data and workflows in ways that competitors cannot easily replicate.

Open-weight models can also increase control over data and deployment. A firm may run the model on its own infrastructure or through a provider of its choice, giving it greater control over where prompts, documents, outputs and logs are processed and retained and whether inputs are used for further training. This control arises from the deployment architecture and its governance, not from openness alone. The advantage is a wider choice of infrastructure rather than exclusive dependence on a provider's standard terms and continued service.

"Open" is not, however, a uniform legal or technical category. Licenses differ in whether they permit commercial use, fine-tuning, redistribution and hosted provision, while many open-weight releases omit training data and development details.[10] Exercising these freedoms can also require more careful model selection, clearer instructions, stronger evaluation, better orchestration and greater internal engineering expertise. Choosing an open-weight model may therefore reduce dependency while increasing the need for organizational knowledge and capacity-building.

9 Frontier labs hold token prices fixed across every venue at which their proprietary models are sold, so hosting on a third-party cloud does not create price competition. Anthropic bills Claude on partner clouds through a "Claude Consumption Unit" (CCU): "One hundred (100) CCU represents $1.00 USD of fees owed for the Services, calculated at the applicable prices on claude.com/pricing#api, after application of any discounts." The identical CCU construct is used on both Claude Platform on AWS (Amazon Bedrock) and Claude in Microsoft Foundry (Azure), so the per-token price is the same whether Claude is bought directly from Anthropic, on Bedrock or on Azure. Anthropic (n.d.), "Pricing," Claude Platform Docs. https://platform.claude.com/docs/en/about-claude/pricing. Accessed 7 October 2026. OpenAI's proprietary GPT models follow the same logic: They were distributed exclusively through Microsoft Azure until the April 2026 restructuring of the Microsoft–OpenAI agreement made the license non-exclusive, after which the models also became available on Amazon Bedrock "at pricing that matches OpenAI's direct rates." In both cases, the models can only be served by licensed partner platforms (Azure, Bedrock, Google Cloud) and cannot be freely hosted by arbitrary inference providers. Markman, J. (2026), "OpenAI Brings GPT-5.5 To AWS Bedrock As Microsoft Exclusive Era Ends," *Forbes*, 6 May. https://www.forbes.com/sites/jonmarkman/2026/05/06/openai-brings-gpt-55-to-aws-bedrock-as-microsoft-exclusive-era-ends/.

10 For example, NVIDIA's Open Model License and OpenMDW-1.1 permit broad commercial use, modification and redistribution of Nemotron 3, while other open-weight licenses impose additional restrictions. Such permissions are separate from technical openness: Access to weights does not imply access to training data, code or full documentation. See NVIDIA (2025), "NVIDIA Nemotron Open Model License"; OpenMDW (n.d.), "OpenMDW-1.1." https://openmdw.ai/license/1-1/.

3 | Competitive advantage from applying AI

Sustained adoption of an AI application depends on two conditions: The application must create value relative to the existing way of performing the task, and the firm must retain enough of that value to cover the costs of implementation and operation and generate an acceptable return. This paper addresses the first condition: value creation.[11] We distinguish this from value capture, that is, how much of the created value is captured by the firm and how much by its suppliers. The application of AI can create economic value even if the adopting firm retains only part of it, for example, because competition passes some of the gains on to customers.

Our first question is therefore whether applying AI improves the economics of performing a particular task. Whether the adopting firm can retain that value is analytically separate. The second condition, which depends on how the value is divided between the adopting firm and its suppliers of models and computing, is the subject of a separate, forthcoming paper.

3.1 | Value creation

The value-creation analysis draws on Michael Porter's framework of competitive advantage.[12] Porter distinguishes two basic types of competitive advantage within an industry. A firm can achieve a lower cost position than its competitors, or it can differentiate itself along dimensions that buyers value and thereby command a higher relative price. Combined with the scope of the market served, these two types underpin Porter's generic strategies: cost leadership, differentiation and focus, with focus taking either a cost or differentiation form. Because this report analyzes the mechanism through which an AI application may create value, rather than the breadth of the market a firm chooses to serve, it concentrates on the two underlying types of advantage: cost advantage and differentiation. Competitive scope remains relevant to a firm's overall strategy, but it does not change the underlying mechanism through which AI contributes to economic value in the use cases examined here.

The distinction provides a simple way to analyze how changes in the performance of an activity translate into economic value. AI may reduce the labor, time or other resources required to perform an activity, contributing to a cost advantage. It may also improve the quality, functionality or delivery of a product or service in ways that buyers value, contributing to differentiation. The two effects can occur together, and their relative importance may differ across use cases.

Throughout our analysis, we therefore assess whether AI creates value by contributing to a cost advantage, differentiation or both and examine the evidence for each mechanism in the use cases considered.

11 These microeconomic effects relate to macroeconomic effects, of course, but we focus only on the net effect of AI uptake within firms. For research on macroeconomic effects linked to productivity growth, see, for instance, Acemoglu, D. (2024), *The Simple Macroeconomics of AI*, NBER Working Paper 32487, National Bureau of Economic Research, May. https://doi.org/10.3386/w32487. On other welfare effects of AI adoption, such as its impact on long-run collective knowledge, see Acemoglu, D., Kong, D. and Ozdaglar, A. (2026), *AI, Human Cognition and Knowledge Collapse*, NBER Working Paper 34910, National Bureau of Economic Research, February. https://doi.org/10.3386/w34910.

12 Porter, M. E. (1985), *Competitive Advantage: Creating and Sustaining Superior Performance*, New York: Free Press.

3.1.1 | Cost advantage

AI can create value by reducing the resources required to produce, deliver or administer a product or service. This may result from shorter processing times, lower coordination and transaction costs, greater output per employee, fewer errors and less rework. AI can also lower the cost of searching for and analyzing information, improve the allocation of resources and reduce the need for external services or additional personnel.

For example, AI coding tools can help software developers complete routine tasks more quickly, document analysis systems can support bank employees in reviewing applications and customer service systems can shorten response times while allowing the same number of employees to handle more inquiries.

A reduction in costs does not necessarily create a lasting competitive advantage. Porter distinguishes operational effectiveness – performing similar activities better – from strategy, which depends on a distinctive position and a coherent system of activities. In his analysis of the internet, Porter argues that, as firms adopt the same widely available technology, "the Internet itself will be neutralized as a source of advantage."[13] Applied to AI, this means that, where competitors can obtain the same tools at a similar cost and deploy them with comparable effectiveness, AI may raise operational effectiveness across an industry without creating a durable difference between firms.

In a separate economy-wide estimate, Nordhaus estimates that U.S. innovators captured only about 2.2% of the social surplus from their innovations between 1948 and 2001, with most of the gains accruing to consumers rather than producers.[14] This historical estimate is not specific to AI, but it illustrates how competition and imitation can limit the share of innovation gains retained by the innovating firm. If competitors can adopt the same technology at a similar price and deploy it with comparable effectiveness, the efficiency gain may therefore quickly become a market standard. Firms may then need to use AI simply to maintain competitive parity, while part of the benefit is passed on to customers through lower prices or better service.

The distribution of this value also has a labor dimension that the firm-level perspective of this paper does not capture. Much of the cost-based value creation discussed in the cases below stems from automating tasks previously performed by people. Acemoglu and Johnson show, for the mechanization of cotton weaving, that such automation can raise average productivity and firm profits while leaving the workers concerned worse off: The price of cloth fell and mill owners earned large profits, but the real wages of handloom weavers more than halved between 1806 and 1820. Whether productivity gains reach workers depends on whether the technology also creates new tasks that raise the marginal productivity of labor and on the bargaining power of workers. Early experimental evidence suggests that generative AI so far substitutes for worker effort more than it complements workers' skills. We therefore treat a cost advantage as value created for the firm and, through competition, its customers, while leaving the effect on workers as a separate question that this paper does not assess.[15]

3.1.2 | Differentiation

AI can create value through differentiation when it improves a product or service in ways that customers notice and value. This may involve higher quality, faster delivery, additional functionality, greater reliability or a better customer experience. In Porter's

13 Porter, M. E. (2001), "Strategy and the Internet," *Harvard Business Review* 79(3), 62–78.

14 Nordhaus, W. D. (2004), *Schumpeterian Profits in the American Economy: Theory and Measurement*, NBER Working Paper 10433, National Bureau of Economic Research. https://doi.org/10.3386/w10433.

15 For example, if AI adoption is not accompanied by new tasks that raise the marginal productivity of labor or sufficient additional hiring in complementary sectors, AI adoption may displace workers while degrading the quality of work of those who remain employed. See Acemoglu, D. and Johnson, S. (2024), Learning from Ricardo and Thompson: Machinery and Labor in the Early Industrial Revolution, and in the Age of AI, NBER Working Paper 32416, National Bureau of Economic Research, May. https://doi.org/10.3386/w32416. See also: "The experimental evidence suggests that ChatGPT largely substitutes for worker effort rather than complementing workers' skills, potentially causing a decrease in demand for workers, with adverse distributional effects as capital owners gain at the expense of workers." Noy, S. and Zhang, W. (2023), "Experimental evidence on the productivity effects of generative artificial intelligence," Science, 381(6654), pp. 187–192. https://doi.org/10.1126/science.adh2586.

terms, differentiation means being unique along dimensions that buyers value.[16] It becomes a competitive advantage when the resulting price premium or equivalent commercial benefit exceeds the additional cost of achieving that differentiation. The economic effect can take several forms: Customers may be willing to pay a higher price, become less sensitive to price increases, choose the product more often, remain with the provider longer or use the service more intensively.

AI can also help firms reach new customer groups or develop entirely new products and revenue streams. For example, a bank may use AI to provide faster and more personalized financial guidance, thereby increasing customer retention, while a software company may add an AI-enabled feature that attracts new users or supports a higher-priced subscription tier. Differentiation therefore creates economic value when improvements that matter to buyers support higher prices, stronger demand, greater retention or new revenue opportunities. It can coexist with lower costs, but its defining mechanism is the effect on buyer value.

The application of AI technologies can contribute to both lower costs and differentiation. A customer service system, for example, may reduce operating costs while also providing faster responses. The analysis therefore identifies the main source of value in each use case while recognizing that both effects may occur together.

The report therefore does not treat AI adoption as intrinsically valuable. For each use case, it identifies the economic task, the principal value-creation mechanism derived from Porter's framework – lower cost, differentiation or both – and, where possible, the capabilities of the model using existing task-specific or industry-specific benchmarks.[17] The AI model is one component of the application. Data, retrieval, tool use, workflow design, testing, governance and integration may be equally important to the realized benefit.

3.2 | Case selection

Independent studies on the economic value created by AI adoption in different sectors remain scarce. Many firms are still experimenting with AI technology, and reported time savings or productivity improvements do not necessarily translate directly into lower costs, higher output or additional revenue. The report therefore focuses on specific use cases and economic activities within selected sectors rather than attempting to estimate the aggregate effect of AI across entire sectors. We selected cases based on three criteria: AI adoption had to be sufficiently advanced to observe concrete applications, the cases had to provide variation in the types of tasks to which AI is applied and relevant model capabilities had to be at least partly assessable using existing benchmarks.

Based on these criteria, we focus on IT and financial services, two sectors in which AI adoption is already comparatively advanced. Within them, we examine three cases: software production, private equity and consumer banking.[18]

Within the financial sector, consumer banking represents high-volume, mass-market and relatively standardized activities. Private equity provides a contrasting case, with lower-volume, document-intensive activities that rely more heavily on judgment. The comparison allows us to examine AI-enabled value creation across two substantially different economic settings within the same broad sector.

Software production provides a third, comparatively structured case. AI adoption is already widespread, and coding tasks can be evaluated more systematically than many less structured activities. We focus on freelance software engineers because productivity gains and differentiation are more directly observable at the individual level. This also helps link observed value creation to model capabilities.

16 Porter, M. E. (1985), *Competitive Advantage: Creating and Sustaining Superior Performance*, New York: Free Press.
17 Porter, M. E. (1985), *Competitive Advantage: Creating and Sustaining Superior Performance*, New York: Free Press.
18 Crane, L., Green, M. and Soto, P. (2025), "Measuring AI Uptake in the Workplace," *FEDS Notes*, Board of Governors of the Federal Reserve System, https://doi.org/10.17016/2380-7172.3724.

The selection of software production is also supported by evidence on current patterns of AI use. Research published by the National Bureau of Economic Research identifies writing and coding among the main current uses of AI.[19] In addition, a survey published by Eurostat in 2025 identified information and communication as the EU sector with the highest share of firms using AI. Across the EU economy, 20% of enterprises with at least 10 employees used AI in 2025, up from 13.5% in 2024. In the information and communication sector, the share was substantially higher, at around 62.5%.[20]

3.2.1 | Software production by freelance software engineers

Within IT, software production is a particularly relevant use case. The use of large language models and coding agents is spreading rapidly among software vendors and IT consulting firms. According to the 2025 Stack Overflow Developer Survey, 84% of surveyed developers were using or planning to use AI tools in their development process, while 47% reported using them daily.[21]

AI is now used across several stages of software development. Applications include generating and explaining code, identifying errors, writing tests, producing documentation and completing multistep development tasks.

Software is therefore a relatively mature AI use case in which adoption is widespread and potential cost effects can be examined at the level of identifiable workflows. It is also suitable for analyzing model requirements because coding performance can be tested more systematically than performance in many less structured business activities. Several mature AI model benchmarks focus on various aspects of software engineering, and we use them to determine whether open-weight models have sufficient capabilities compared with proprietary models.

3.2.2 | Consumer-facing banking

Consumer banking provides a high-volume and relatively standardized financial services case in which AI adoption is already widespread.

The European Banking Authority found that 55% of surveyed banks in 2025 were already using general-purpose or agentic AI, particularly in consumer-facing processes.[22] Reported applications included digital assistants, fraud detection, support for customer service employees and the automated provision of information to customers.

3.2.3 | Private equity

Private equity provides a contrasting financial services case, characterized by lower-volume, document-intensive workflows that rely more heavily on judgment. AI systems are increasingly being developed for relevant tasks. Anthropic, for example, has launched a dedicated financial services offering for workflows such as due diligence, market research, portfolio analysis and financial modeling. At the same time, benchmarks such as Vals AI's Finance Agent Benchmark evaluate models on realistic financial research tasks.[23,24]

Private equity is therefore useful for assessing whether AI can create value in workflows that combine information retrieval, document analysis and judgment and for examining how far relevant model capabilities can be evaluated systematically.

19 Chatterji, A. et al. (2025), *How People Use ChatGPT*, NBER Working Paper 34255, National Bureau of Economic Research. https://www.nber.org/papers/w34255.

20 Eurostat (2025), "Use of artificial intelligence in enterprises." https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/106920.pdf.

21 Stack Overflow (2025), "2025 Stack Overflow Developer Survey." https://survey.stackoverflow.co/2025.

22 European Banking Authority reports that 92% of EU banks were deploying AI and 55% of surveyed banks were using general-purpose or agentic AI in consumer-facing processes. See European Banking Authority (2025), *Rising application of AI in EU banking and payments sector*, Digital Finance Factsheet Series, 25 September. https://www.eba.europa.eu/sites/default/files/2025-09/146b3558-d026-47bf-a872-f05e93ed30d2/Rising%20application%20of%20AI%20in%20EU%20banking%20and%20payments%20sector.pdf. For an example of customer-facing applications, see BBVA (2024), "What AI algorithms does BBVA use to boost its customers' finances?", 2 August. https://www.bbva.com/en/innovation/what-ai-algorithms-does-bbva-use-to-boost-its-customers-finances/

23 Lin, B. (2026), "Anthropic Releases New AI Agents for Financial Services Firms," *The Wall Street Journal*, 5 May, https://www.wsj.com/tech/ai/anthropic-releases-new-ai-agents-for-financial-services-firms-e2829b37.

24 Vals AI (n.d.), "Finance Agent Benchmark." https://www.vals.ai/home.

3.3 | Methodology

3.3.1 | Research design

The report uses an exploratory qualitative multiple case study design. The analysis focuses on specific use cases and economic activities within the three cases and combines desk research with semi-structured interviews. Desk research and semi-structured interviews are used to identify where AI is applied and whether it can reduce the resources required for an economic activity or increase the value of the resulting product or service.

Benchmarks provide proxies for specific model capabilities relevant to the economic activities identified. We use SWE-rebench for software engineering tasks and for financial services we use τ^3-Banking Benchmark and Vals Finance Agent Benchmark.[25][26][27] We compare benchmark performance to assess whether these capabilities are available in open-weight models.

3.3.2 | Semi-structured interviews

We conducted a total of 12 interviews. Freelance software engineers were selected from our pre-existing network and through referrals from interviewees. For consumer banking, we conducted background interviews, but the analysis relies primarily on desk research because many of the AI applications identified were still being evaluated rather than used in day-to-day operations. Within private equity, we conducted interviews with senior analysts and portfolio managers.

Interviews followed a light semi-structured guide and were recorded where permitted. For the software engineers, we focused on senior developers with an existing customer base and at least 10 years of experience building software applications. The organizations interviewed in financial services represented large, well-established actors.

3.3.3 | Limitations

The analysis is necessarily selective and does not provide a comprehensive account of AI adoption and value creation across the sectors examined. Detailed public evidence on AI adoption, associated costs and realized economic benefits remains limited and is often based on company-reported examples. Many organizations we contacted were also still evaluating AI applications or were unwilling to disclose their strategies and operational data. Within financial services, interviewees did not provide detailed operational metrics, such as annual AI spending or quantified benefits.

Benchmarks provide proxies for capabilities relevant to particular tasks, but benchmark performance does not establish that a model will perform sufficiently well in a specific workflow or that its use will create economic value. The realized outcome also depends on the wider application, including data, tools, workflow design and integration. The prices reported in Chapter 5 are snapshots at the time of writing and are volatile.

Given the limited number of interviews, the findings should be interpreted as illustrative rather than representative and cannot be generalized to the sectors as a whole. The interview sample is also likely to be weighted toward relatively experienced or early adopters of AI. The combination of case evidence, interviews and benchmarks is therefore best suited to identifying and assessing concrete mechanisms of AI-enabled value creation. It does not support conclusions about how prevalent those mechanisms are or about their aggregate economic effects across the sectors examined.

25 SWE-rebench (n.d.), "SWE-rebench v2 Leaderboard." https://swe-rebench.com/.
26 Artificial Analysis (n.d.), "τ^3-Banking Benchmark Leaderboard." https://artificialanalysis.ai/evaluations/tau3-banking.
27 See footnote 24.

4 | Applying the framework to three AI use cases: empirical analysis

This chapter focuses first on how AI creates value in the three cases, irrespective of the model type used. It then compares open-weight and proprietary models on benchmarks relevant to the activities examined. Questions of lock-in, token prices, supplier rents and the distribution of gains are addressed in Chapters 5 to 7, while the dynamics of value capture are examined further in a forthcoming paper. Following Porter's framework, each use case is evaluated according to whether AI contributes to cost advantage, differentiation or both and which capabilities are required to realize these benefits.

The evidence indicates that AI can create value in the cases examined, but not uniformly. The strongest qualitative evidence concerns productivity in software production and recurring workflows in fund management. Consumer-facing banking provides little evidence that AI is separately monetized or currently supports durable differentiation.

4.1 | Software production

Software production is a comparatively mature use case for analyzing AI-enabled value creation. AI tools are already used in operational workflows, and their outputs can be assessed at the level of specific tasks. Here, software production covers the design, development, testing, documentation, deployment and maintenance of software applications and systems.

The strongest evidence concerns productivity. Interviewees consistently reported that AI reduced the time required for specific development tasks, supporting a qualitative case for value creation through cost advantage. The evidence for differentiation is weaker: Faster delivery appears increasingly to be expected by customers rather than rewarded through a separate premium.

It is worth noting that at the time of the interviews, none of the developers interviewed used open-weight models. They were aware of their existence and considered them a viable alternative. However, all interviewees said that current subscription offerings from proprietary AI labs were inexpensive enough that switching to an open harness using pay-per-use open-weight models – and rebuilding their workflows around it – was not worth the required investment. While open-weight models may be cheaper on a per-task basis, subscriptions offer cost certainty and, according to the interviewees, enable value creation that substantially exceeds the monthly subscription fee for model access.

4.1.1 | Value creation

4.1.1.1 | Cost advantage

In software production, AI can lower costs by reducing the developer time or other resources required to produce a given output. This can lower labor costs, increase the volume of work completed with a fixed workforce or shorten delivery times.

The interviews provide consistent evidence of this effect. Interviewees reported using AI to accelerate routine coding, analyze unfamiliar code, assist with debugging and testing and reduce the time required to document or revise a solution. All freelance software developers interviewed reported that AI-assisted development increased their productivity, allowing them to complete tasks more quickly and work on more projects in parallel. This finding is consistent with other studies reporting self-assessed productivity gains among technical workers using AI tools.[28]

28 Becker, J. (2026), "Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity," METR, 11 May. https://metr.org/blog/2026-05-11-ai-usage-survey/#summary.

Some interviewees also reported greater context-switching fatigue, which may offset part of the gain, but the reported net effect remained positive.

At the firm level, these productivity gains can reduce the labor hours required to deliver a given amount of work, allow more projects to be completed with the same workforce or shorten time to market. For independent developers and consultancies, they may also increase the number of clients or projects that can be served within a given period.

4.1.1.2 | Differentiation

In software production services, differentiation would arise if AI-enabled improvements affected what customers pay, which provider they choose or how much work they commission. Relevant indicators include higher rates, a premium for faster or more reliable delivery and additional demand for specialized capabilities. The available interview evidence relates primarily to development services rather than software products.

Interviewees did not report higher rates, a delivery premium or another observable demand effect attributable to AI. Several instead indicated that customers increasingly expect faster results because AI-assisted development tools are widely available. Faster delivery may therefore benefit the customer without distinguishing one developer from another. This suggests that higher speed may be becoming a market requirement rather than a durable source of differentiation.

Differentiation may still be possible in projects where speed, reliability or specialized technical capabilities are particularly important. The ability to design and manage complex AI-enabled development workflows may also become a differentiating skill. The interviews do not yet show that these capabilities command higher prices or generate additional demand.

4.1.2 | Capabilities

The productivity gains described above depend on models being able to perform more than isolated code generation tasks. Developers use AI to understand existing code, identify errors, implement changes, write tests and work across larger codebases. To support these workflows, a model must therefore be able to understand an existing repository, identify the relevant files and dependencies, translate a request into code changes and use tools such as test suites or terminals to verify whether the solution works.

We use SWE-rebench v2 to assess these capabilities. It evaluates models on real software engineering issues drawn from recent GitHub repositories. Models have to understand an issue within an existing codebase and implement a working fix. SWE-rebench also regularly introduces new issues, reducing the risk that benchmark performance reflects prior exposure to the underlying tasks.[29]

Table 1 | SWE-rebench v2 scores for leading proprietary and open-weight models

Proprietary model	Issues resolved	Open-weight model	Issues resolved
Fable 5	64.5%	GLM-5.2	62.9%
Grok 4.5	63.8%	MiniMax M3	47.2%
Opus 5	63.4%	MiMo V2.5 Pro	46.5%

Source: SWE-rebench (n.d.), "SWE-rebench v2 Leaderboard." https://swe-rebench.com/. | BertelsmannStiftung

SWE-rebench reports the average share of software engineering issues successfully resolved across five independent runs of the benchmark. In the current evaluation window, this covers 111 real-world issues from 65 GitHub repositories. The scores reported here reflect the leaderboard as of Sept. 4, 2026.

29 Badertdinov, I. et al. (2025), "SWE-rebench: An Automated Pipeline for Task Collection and Decontaminated Evaluation of Software Engineering Agents," *NeurIPS 2025*, arXiv:2505.20411. https://doi.org/10.48550/arXiv.2505.20411.

The strongest open-weight model performs very close to the leading proprietary models. GLM-5.2 resolves 62.9% of the issues, compared with 64.5% for Fable 5, 63.8% for Grok 4.5 and 63.4% for Opus 5. The gap between the strongest open-weight model and the benchmark leader is therefore only 1.6 percentage points.[30]

The benchmark therefore suggests that leading open-weight models already provide software engineering capabilities close to those of proprietary frontier models. This does not imply identical reliability in practice, and even the strongest models still fail a meaningful share of benchmark tasks. In software development, however, models are typically used within workflows in which experienced developers review and correct their output. A small remaining performance gap may therefore have limited practical relevance for many coding, debugging and repository-level tasks. Access to a proprietary frontier model is thus no longer a prerequisite for obtaining the capabilities needed to generate the productivity gains described above.

4.2 | Consumer-facing banking services

In consumer banking, AI is increasingly integrated into digital assistants, customer support, spending analysis and the provision of routine information. These high-volume and relatively standardized services provide a useful setting in which to assess both operational efficiency and customer-facing differentiation.

The evidence supports a plausible case for operational efficiency, but does not establish the financial effect of those efficiency gains. The services reviewed also provide little evidence that customer-facing AI currently supports a separate price premium or durable differentiation.

4.2.1 | Value creation

4.2.1.1 | Cost advantage

In consumer-facing banking, AI can lower costs when it reduces the employee time or other resources required to answer inquiries, provide routine information or complete recurring service tasks. This can lower the cost per interaction, allow a fixed workforce to handle more contacts or shorten response and resolution times.

The banks reviewed report substantial use of automated service channels.[31] Postbank stated that its virtual assistant "Paula" handled customer inquiries that would otherwise have resulted in around 25% of calls to its service center.[32] ING reported 85,000 weekly contacts through its chatbot, while its preceding bot resolved 40 to 45% of customer inquiries without further intervention.[33] ABN AMRO reported approximately 3.5 million annual text and voice conversations through its virtual assistant "Anna," while Société Générale reported around 4 million annual conversations through "Eliott," with more than 70% handled end to end, as well as approximately 6,000 daily interactions through its "SoBot" chatbot.[34,35] Intesa Sanpaolo and bunq also reported high shares of requests or support activity

30 SWE-rebench (n.d.), "SWE-rebench v2 Leaderboard." https://swe-rebench.com/.

31 European Banking Authority (2025), *Rising application of AI in EU banking and payments sector.* The EBA reports that 92% of EU banks deploy AI and that 55% of surveyed banks use general-purpose or agentic AI in consumer-facing processes. For examples of consumer-facing applications, see BBVA (2024), "What AI algorithms does BBVA use to boost its customers' finances?"; CaixaBank (2025), "CaixaBank develops an AI agent to help customers explore products in the app."

32 Deutsche Bank (n.d.), "Technology Transformation." https://www.db.com/what-we-do/focus-topics/tech/. Accessed 7 June 2026.

33 McKinsey & Company (2024), "Banking on innovation: How ING uses generative AI to put people first," https://www.mckinsey.com/industries/financial-services/how-we-help-clients/banking-on-innovation-how-ing-uses-generative-ai-to-put-people-first.

34 Microsoft (2024), "ABN AMRO Bank migrates to Microsoft Copilot Studio to support over 3.5 million customer conversations," Microsoft Customer Stories. https://www.microsoft.com/en/customers/story/19754-abn-amro-bank-microsoft-copilot-studio.

35 Societe Generale (2024), "Societe Generale shares its strategy for integrating generative AI into the financial sector," 10 October. https://www.societegenerale.com/en/news/all-news/societe-generale-shares-its-strategy-integrating-generative-ai-financial-sector.

handled through their assistants, although these figures are self-reported and not directly comparable.[36]

These volumes indicate that banks may save substantial employee time by automating routine interactions. At the firm level, this could increase service capacity and allow employees to concentrate on cases requiring judgment.

However, the net cost effect depends on the quality of the automated service. If assistants provide inadequate answers, fail to resolve requests or require customers to repeat information when a case is escalated, the resulting follow-up and correction work may offset part of the initial time savings. Poor service may also increase complaints, repeated contacts or customer attrition, creating additional costs elsewhere in the organization. If customers perceive AI-enabled service as lower quality, a positive cost effect could also be partly or fully offset by a negative differentiation effect.

4.2.1.2 | Differentiation

For customer-facing AI, the relevant question is whether an improved service affects customers' willingness to pay, choice of provider, retention or use. The existence of an assistant or personalized guidance alone does not establish such an effect.

The following comparison examines how selected banks and neobanks position and price their customer-facing AI features. It focuses on whether the feature is charged for separately, restricted to a premium tier or marketed as a reason for paying more.

36 Intesa Sanpaolo (2023), "Intesa Sanpaolo only Italian bank selected at the Qorus-Accenture Banking Innovation Awards," press release, 4 December. https://group.intesasanpaolo.com/en/newsroom/press-releases/2023/12/intesa-sanpaolo-only-italian-bank-selected--at-the-qorus-accentu.

Table 2 | Customer-facing AI features, account pricing and premium-tier benefits at selected banks and neobanks

Institution	Customer-facing AI feature	AI pricing treatment	Basis of premium tier	Source(s)
BBVA, Spain	"Blue" natural-language assistant; financial coach; real-time transaction categorization; expected-expense prediction; planned 2026 ChatGPT application	Bundled in the application; free digital account available	No separate premium AI feature identified	BBVA
CaixaBank, Spain	Generative-AI agent recommending and comparing cards; purchasing agent	Bundled; account fees vary	No separate premium AI feature identified	CaixaBank agent; CaixaBank purchasing agent
Deutsche Bank/Post-bank, Germany	"Paula" customer-service chatbot	Bundled; fee-based account	No separate premium AI feature identified	UC Strategies
ING, Netherlands/ Germany	Generative-AI customer-service chatbot	Bundled; German account free with at least € 1,000 monthly inflow, otherwise € 4.90 per month	The reported fee change is unrelated to AI	McKinsey; Pro Finance
ABN AMRO, Netherlands	"Anna" text and voice assistant	Bundled; fee-based account	No separate premium AI feature identified	ABN AMRO
Intesa Sanpaolo, Italy	"Ellis" assistant; isybank digital service	Bundled; fee-based account, with low- or zero-fee isybank options	No separate premium AI feature identified	Qorus
Société Générale, France	"Eliott" callbot at BoursoBank; "SoBot" chatbot	Bundled; Société Générale accounts are fee-based, while BoursoBank offers a free account	No separate premium AI feature identified	Société Générale Sobot; Société Générale Eliott
BNP Paribas, France	Customer experience personalization and generative AI chat among more than 750 production AI use cases	Bundled; fee-based account	No separate premium AI feature identified	BNP Paribas
N26, Germany/ EU	"Insights" transaction categorization, subscription overview and monthly wrap-up; support chatbot	Available on the free Standard account	Premium tiers (€ 4.90, € 9.90 and € 16.90) add cards, insurance and ATM or foreign-exchange benefits rather than a distinct AI service	N26
Revolut, Lithuania/ EU	"AIR" assistant for spending analysis, card controls, subscriptions and eSIM services	Bundled across tiers; Standard account is free	Paid tiers are marketed through travel, insurance, subscriptions, eSIM data and savings benefits rather than AI	Revolut
bunq, Netherlands/ EU	"Finn" conversational assistant, including spending queries, translation and receipt scanning	Available across paid plans; the lowest-priced account is approximately € 3.99 per month	Higher-priced plans (€ 9.99 and € 18.99) add sub-accounts, insurance, eSIM data, transfers and foreign-exchange benefits rather than a distinct AI service	bunq
Monzo, UK	AI support assistant; "Trends" categorization; fraud and machine learning features	AI is not separately monetized; max is approximately € 20 per month	Travel and phone insurance, breakdown cover, savings benefits and other services	Monzo
Starling, UK	Spending Intelligence, an agentic assistant and Scam Intelligence	Free and opt-in; the bank reportedly has no plans to charge for the features	No paid personal-account tier	Starling

Source: own table

Across the sample, customer-facing AI is generally bundled into existing accounts or available across tiers without a separate charge.[37] Where banks and neobanks offer premium accounts, the marketed benefits are mainly conventional features such as insurance, travel benefits, additional cards, ATM or foreign exchange allowances, subscriptions and higher savings rates.[38] The evidence does not show that AI itself supports a premium price.[39] As similar capabilities become more widely available, these features appear more likely to become a competitive requirement than a durable basis for differentiation.[40]

Conversely, human customer support could itself become a source of differentiation that customers perceive as more valuable, particularly if AI-powered chatbots provide lower-quality service than human agents.

AI implementation may nevertheless differentiate banks. Proprietary customer data, distribution, scale and the quality of integration may allow some incumbents to provide a more useful service than competitors using similar underlying models.[41] The available evidence, however, does not establish that these differences currently increase customers' willingness to pay or produce a sustained premium.

Fintech startups are a partial exception. Alongside AI enhancements to existing services, some have built genuinely new AI-native product propositions that open new revenue streams.[42] However, because these firms compete using the same broadly available technology, the prices of these new services remain relatively low.[43] This implies either that the incremental value to customers remains modest or that providers have not yet been able to convert that value into a price premium.

4.2.2 | Capabilities

The operational gains described above depend on models being able to handle routine customer interactions accurately and with limited employee involvement. Customer-facing assistants must understand a customer's request, retrieve the relevant information from product and policy documentation, apply that information to the customer's circum-

37 N26 "Insights" is available on the free €0 Standard tier; Revolut's "AIR" assistant rolled out to UK customers with no separate price. See Sacra (n.d.), "N26 revenue, valuation & funding." https://sacra.com/c/n26/. Accessed 7 October 2026; Revolut (2026), "Revolut enters new era of money intelligence with launch of AI assistant," 9 April. https://www.revolut.com/news/revolut_enters_new_era_of_money_intelligence_with_launch_of_ai_assistant/.

38 Revolut's subscription revenue reached approximately US$936 million in 2025, up roughly 67% year on year, driven by uptake of its paid account tiers and their associated perks rather than by AI-specific features. See Revolut (2026), "Revolut reports record profit of $2.3bn for 2025 as revenue surges to $6bn," 24 March. https://www.revolut.com/news/revolut_reports_record_profit_of_2_3bn_for_2025_as_revenue_surges_to_6bn/.

39 Banque de France, *Observatoire des tarifs bancaires*, reported bank service prices up 2.7% between February 2025 and February 2026; German banks raising account fees in 2024 cited ECB low-rate policy, not AI. See Lilley, Z. (2026), "French bank account fees increased 2.7% between 2025 and 2026," *The Connexion*, 2 July. https://www.connexionfrance.com/practical/french-bank-account-fees-increased-27-between-2025-and-2026/800119; mutual (2020), "Deutsche Bank erhöht Gebühren für Girokonten," 1 July, updated 13 August 2025. https://www.mutual.de/konto/deutsche-bank-erhoeht-gebuehren-fuer-girokonten/.

40 Forrester argues that an AI strategy focused mainly on cheaper, faster servicing accelerates commoditization rather than building competitive advantage. See Clarke, A. (2026), "AI Isn't Fixing Retail Banking's Customer Growth Problem — It's Exacerbating It," Forrester, 14 May. https://www.forrester.com/blogs/ai-isnt-fixing-retail-bankings-customer-growth-problem-its-exacerbating-it/.

41 McKinsey & Company (2026), "From AI table stakes to AI advantage: Building competitive moats," *McKinsey Quarterly*, 15 May. https://www.mckinsey.com/capabilities/quantumblack/our-insights/from-ai-table-stakes-to-ai-advantage-building-competitive-moats. When every competitor has access to the same AI models, durable advantage comes from proprietary data, scale and execution rather than model access. On the scale of the prize, see Sachse, H. et al. (2025), *From Branches to Bots: Will AI Agents Transform Retail Banking?*, Boston Consulting Group, 5 November, projecting more than $370 billion in annual retail banking profit potential by 2030. https://www.bcg.com/publications/2025/branches-to-bots-will-ai-transform-retail-banking.

42 Cleo pairs an AI-chat interface with genuinely new propositions: cash advances, automated subscription cancellation/negotiation and a credit-builder card. See Cleo AI (n.d.), "Cleo AI: Cash Advance & Money," Google Play. https://play.google.com/store/apps/details?id=com.meetcleo.cleo&hl=en_US. Accessed 14 July 2026.

43 European neobank premium tiers remain low (N26 €4.90/€9.90/€16.90; Revolut Plus/Premium/Metal ≈ €2.99–€45; bunq €3.99–€18.99), and standalone AI-PFM apps such as Cleo run US$5.99–$14.99/month. See Sacra (n.d.), "N26 revenue, valuation & funding"; Revolut (2026), "Ask, and AIR makes it happen," 9 April. https://www.revolut.com/blog/post/air-ai-by-revolut/; Cleo AI (n.d.), "Cleo AI: Cash Advance & Money."

stances and maintain context across the interaction. For more advanced services, they may also need to use account information or tools to complete a request. These capabilities determine whether an AI assistant can resolve an interaction without employee intervention and thereby reduce the employee time required for customer service.

To assess how open-weight models compare with proprietary models on these capabilities, we use τ³-Banking as the main benchmark for customer service tasks. It evaluates AI agents in a simulated customer service environment containing 698 interconnected documents across 21 banking product categories. The 97 tasks cover workflows such as account and card recommendations, disputes, account freezes, replacement cards and product changes. Models must locate the relevant documentation, interpret the applicable rules and complete the required actions through a sequence of tool calls. The benchmark therefore combines knowledge retrieval, policy application, customer interaction and tool use in a way that closely reflects the consumer banking workflows examined above.

The score measures the average share of banking scenarios completed successfully across repeated runs. A task is counted as successful only when the interaction produces the required final state in the simulated bank database. This may require the model to retrieve several relevant documents, apply their rules correctly and complete the full sequence of account actions. The score therefore measures end-to-end resolution rather than the quality of an isolated answer.[44]

Table 3 | τ³-Banking scores for leading proprietary and open-weight models

Proprietary model	Tasks completed	Open-weight model	Tasks completed
Qwen3.8 Max	51.3%	GLM-5.3 (max)	50.3%
Grok 4.6 (high)	50.7%	Qwen3.8 2.4T A95B	49.1%
Muse Spark 1.3 (max)	50.5%	Qwen3.8 27B (xhigh)	48.0%

Source: Artificial Analysis (n.d.), "τ³-Banking Benchmark Leaderboard." https://artificialanalysis.ai/evaluations/tau3-banking.

BertelsmannStiftung

The leading open-weight models perform very close to the leading proprietary frontier models shown. GLM-5.3 completes 50.3% of the tasks, compared with 51.3% for the benchmark leader, a difference of 1 percentage point. It is only 0.4 percentage points behind Grok 4.6 and 0.2 percentage points behind Muse Spark 1.3 (max). Two further open-weight models reach 49.1% and 48.0%, placing three open-weight models among the six highest-scoring systems shown on the benchmark leaderboard.[45] The scores reported here reflect the leaderboard as of Sept. 10, 2026.

These results suggest that leading open-weight models can provide customer service capabilities close to those of the strongest proprietary models on this benchmark. No model resolves every case, so deployment still requires bank-specific documentation, access controls, verification procedures and escalation to employees. However, these requirements apply to proprietary and open-weight models alike. For the routine workflows tested here, the benchmark provides no indication that access to a proprietary frontier model is necessary to obtain strong underlying model capabilities.

The banks examined rely predominantly on proprietary models, which limits the direct evidence on open-weight deployment in consumer banking. Our assessment of open-weight capabilities therefore draws mainly on public benchmarks. Revolut's PRAGMA provides a separate example of specialized model development using a bank's own data.

44 Artificial Analysis (n.d.), "τ³-Banking Benchmark Leaderboard." https://artificialanalysis.ai/evaluations/tau3-banking.
45 Artificial Analysis (n.d.), "τ³-Banking Benchmark Leaderboard." https://artificialanalysis.ai/evaluations/tau3-banking.

Revolut's PRAGMA model provides a contrasting example of specialized model development at relatively modest cost, with measurable improvements over existing proprietary AI models. Revolut published its process and findings in a research paper that explicitly sets out the tasks a model must perform in this setting and the capabilities those tasks require.[46] PRAGMA is a family of foundation models for banking event sequences rather than language. It treats a customer's history as a sequence of discrete, time-stamped, variable-length records and is pretrained on approximately 24 billion events, or some 207 billion tokens, drawn from around 26 million customers in 111 countries over 25 months. It is evaluated on economically significant tasks in consumer banking: credit scoring, external fraud detection, product recommendation, customer communication targeting, recurrent transaction detection and lifetime value prediction. Compared with Revolut's own production models, PRAGMA reports relative improvements of 64.7% in fraud detection recall, with 16.7% higher precision, 130.2% in credit scoring and 40.5% in product recommendation.[47]

The capabilities these tasks require are not those for which frontier language models are optimized. Revolut's team explicitly rejected the apparent shortcut of serializing banking records as text and feeding them to a standard language model. Fraud detection and credit risk are sequence modeling and classification problems involving proprietary behavioral data, not generative language problems.

Moreover, the resources required to create PRAGMA were modest. The three variants, with 10 million, 100 million and 1 billion parameters, were trained on rented clusters of 16 to 32 NVIDIA H100 GPUs and converged within two weeks.[48] This level of computing capacity may be within reach of a midsize European firm. However, it depends on the firm's own data and engineering capabilities rather than access to frontier models. NVIDIA's Llama-Nemotron-Embed-1B-v2, released under the NVIDIA Open Model License, served as a supporting open-weight model and was used to embed merchant descriptions and other unstructured text fields.[49] PRAGMA is therefore an example of a bank creating value from its own data by engineering AI around a specific set of economically relevant tasks.

46 Ostroukhov, M. et al. (2026), "PRAGMA: Revolut Foundation Model," arXiv preprint 2604.08649, https://arxiv.org/abs/2604.08649.

47 See footnote 46.

48 PRAGMA-S (10M parameters) and PRAGMA-M (100M) were trained on 16 NVIDIA H100 GPUs, while PRAGMA-L (1B) was trained on 32, with convergence in approximately two days for the smallest and two weeks for the two larger variants. NVIDIA's case study specifies that training ran on H100 GPUs rented from Nebius AI Cloud, an EU-based provider, alongside NVIDIA cuDF for data processing, and reports two to five times higher training throughput than padded baselines through sequence packing and dynamic batching. Downstream tasks are addressed either by freezing the backbone and training a lightweight head on the extracted embeddings or by LoRA fine-tuning of the QKV projections and MLP layers (rank 8, alpha 8) at a parameter overhead of 2%–4%. See Ostroukhov et al. (2026), *PRAGMA: Revolut Foundation Model;* NVIDIA (n.d.), "Revolut Builds a Transaction Foundation Model With NVIDIA Accelerated Computing Platform," case study. https://www.nvidia.com/en-us/case-studies/revolut/. Accessed 7 October 2026; Nebius (n.d.), "Revolut on the Inference Frontier," case study. https://nebius.com/customer-stories/revolut. Accessed 7 October 2026.

49 NVIDIA's Llama-Nemotron-Embed-1B-v2, an open-weight text-embedding model released under the NVIDIA Open Model License (with the Llama 3.2 Community License Agreement applying additionally), is used to embed merchant descriptions and other unstructured text fields. See NVIDIA (n.d.), "llama-nemotron-embed-1b-v2," model card, Hugging Face. https://huggingface.co/nvidia/llama-nemotron-embed-1b-v2. Accessed 7 October 2026.

4.3 | Private equity

Fund management covers two related but distinct channels. The first is the use of AI within the fund manager's own operations and investment process, including compliance, annual reporting, knowledge sharing, market research, company screening and due diligence. The second is the use of AI as part of a value-creation plan in portfolio companies. The analysis treats internal fund workflows as the primary unit of analysis and considers portfolio company adoption as a separate channel for which the evidence is more limited, although interviewees saw potentially greater value there.

4.3.1 | Value creation

4.3.1.1 | Cost advantage

In fund management, AI can lower costs when it reduces the professional time or other resources required to perform recurring administrative and investment analysis tasks. This may lower the cost of producing a given analytical output, allow a fixed team to assess more opportunities or shorten the duration of reporting, screening and diligence processes.

The private equity interviews provide clearer evidence of this effect than the consumer banking material. Funds are applying AI to compliance, annual reporting and knowledge sharing, as well as to market research, company screening, financial model preparation and due diligence. Interviewees reported time savings in compliance and screening, less repetitive analysis and shorter review processes where AI reduced the time required to assemble and assess information. The strongest evidence concerns recurring, document-heavy tasks.

The interviewees indicated that the objective is generally not to reduce headcount. Funds instead seek to shift investment professionals away from document screening and repetitive modeling toward judgment-intensive work, portfolio company support and relationship management. At the firm level, the gross benefit is therefore higher effective capacity and shorter process duration rather than a directly observed reduction in payroll. No fund provided operational or financial metrics, so the magnitude of the effect remains qualitative.

4.3.1.2 | Differentiation

In fund management, differentiation may arise where AI improves investment performance, accelerates capital deployment or makes the fund a more attractive counterparty. These effects differ from cost advantage because they depend on better investment outcomes or a stronger market position rather than simply on reducing professional hours.

Earlier identification of potential targets may allow a fund to enter a process sooner, while more effective assessment may improve investment selection. Faster screening and due diligence may also shorten the period for which committed capital remains uninvested. This can be important because many funds hold uncalled capital, or so-called dry powder.[50] This is capital committed by investors but not yet invested, potentially because of extensive due diligence, capacity constraints or lengthy search processes. Shorter and more reliable processes may additionally strengthen a fund's position with sellers and other counterparties. The interviews identified these mechanisms but did not provide quantitative evidence of their effect on fund performance. The available evidence is therefore indicative rather than conclusive.

Interviewees considered the larger opportunity to lie in portfolio companies. AI can form part of a value-creation plan by improving products or operations during the holding period. This may ultimately affect fund returns, but the available material does not provide examples or metrics linking a specific implementation to portfolio company performance. The principal finding is therefore that funds see substantial potential but continue to face difficulties turning it into practical applications.

4.3.2 | Capabilities

The productivity gains described above depend on models being able to work across document-heavy financial workflows. For compliance and reporting,

50 Simlai, P. E. (2025), "Financial environment, dry powder, and the dynamics of private equity valuations," *International Review of Economics & Finance*, 104, article 104761. https://doi.org/10.1016/j.iref.2025.104761.

models must locate relevant information, synthesize it across documents and produce consistent and traceable outputs. Market research, company screening, financial model preparation and due diligence also require models to identify relevant filings and disclosures, reconcile information across sources, apply financial conventions, perform calculations and compare companies or transactions. These capabilities reduce the repetitive search, screening and modeling work that interviewees identified as the main source of time savings.

We use Finance Agent v2, the current version of the Finance Agent Benchmark, to assess these capabilities. It evaluates models on work typically performed by an entry-level financial analyst using public company filings. The current test set contains 450 held-out questions across nine categories, including qualitative and quantitative analysis, market analysis, comparables, precedent transactions, financial adjustments, earnings and disclosure analysis and financial modeling. Models use the same agent setup and have access to SEC filings, web search, document retrieval tools, a calculator and price histories.[51] The benchmark therefore closely reflects the multistep research, analysis and modeling required for company screening and parts of due diligence, although it does not cover final investment judgment or work with private deal materials.

Table 4 | Finance Agent v2 scores for leading proprietary and open-weight models

Proprietary model	Finance Agent v2 score	Open-weight model	Finance Agent v2 score
Gemini 3.8 Flash	61.44%	GLM 5.3 Flash	57.85%
Muse Spark 1.2	60.60%	GLM 5.3	55.84%
Gemini 3.7 Flash	59.04%	Kimi K3	54.36%

Source: Vals AI (n.d.), "Finance Agent Benchmark." https://www.vals.ai/home.

The score measures performance across several requirements within each question rather than only whether the final answer is entirely correct. Each question contains weighted checks covering the relevant facts, calculations and reasoning steps. Some are classified as essential, meaning that a response receives no credit for the question if it misses a load-bearing fact or number. Models otherwise receive partial credit for the requirements they satisfy. Each model is evaluated three times on the full test set, and the reported score is the average across these runs.[52]

The leading open-weight model scores within 3.59 percentage points of the three leading proprietary models shown. GLM 5.3 Flash scores 57.85%, compared with 59.04% for Gemini 3.7 Flash. It is 2.75 percentage points behind Muse Spark 1.2 and 3.59 percentage points behind the benchmark leader. GLM 5.3 and Kimi K3 also achieve scores close to those of the proprietary frontier models. The results therefore show that leading open-weight models already provide strong capabilities for financial research, company analysis and modeling. The scores reported here reflect the leaderboard as of Sept. 4, 2026.

In private equity, these models are mainly used to support analysts rather than make investment decisions autonomously. They can accelerate research, screening, financial analysis and parts of due diligence, while investment professionals remain responsible for checking the underlying evidence, assumptions and calculations. In this setting, an open-weight model does not need to match the best proprietary model on every task to create value. The benchmark results show that leading open-weight models provide strong financial research capabilities relevant to these activities. Proprietary frontier models are therefore not the only source of the capabilities that firms can draw on for financial research and analysis.

51 Vals AI (n.d.), "Finance Agent Benchmark." https://www.vals.ai/home.
52 Vals AI (n.d.), "Finance Agent Benchmark." https://www.vals.ai/home.

4.4 | Cross-case conclusion

Across the three use cases, the evidence for AI-enabled value creation is strongest through cost advantage. In software production, interviewees reported that AI reduced the developer time required for coding, debugging and related tasks. In private equity, the main benefits likewise came from time savings in research, screening and other document-intensive workflows. In consumer banking, automated service channels can handle substantial volumes of routine interactions and therefore offer a plausible route to lower operating costs, although the resulting financial effect is not yet established.

The evidence for differentiation is weaker. In software production, faster delivery increasingly appears to be expected by customers rather than rewarded through higher prices. In consumer banking, AI-enabled services are generally bundled into existing accounts and are not separately monetized. In private equity, AI may improve investment selection, accelerate capital deployment or strengthen a fund's position with counterparties, but the available evidence does not yet establish these effects. Across the three cases, AI therefore currently creates value more clearly through improvements in efficiency and capacity than through differentiated products or services.

The separate benchmark comparisons show that leading open-weight models perform close to proprietary models on repository-level software engineering, customer service in banking and financial analysis.

The gap to the benchmark leader is 1.6 percentage points in software engineering, 1 percentage point in consumer banking and 3.59 percentage points in financial analysis. The results therefore suggest that the model capabilities required for many of the workflows examined in this report are no longer confined to proprietary frontier models. Firms do not necessarily need access to the most capable proprietary model to obtain strong capabilities for these tasks.

The narrowing capability gap makes price increasingly important to model choice. Where open-weight and proprietary models provide broadly comparable capabilities for a given workflow, the price of accessing those capabilities becomes more important. Chapter 5 therefore compares the prices of proprietary and open-weight models and examines whether open-weight models can provide comparable capabilities at lower prices.

5 | Prices, capability and barriers to open-weight adoption

Chapter 4 identified how AI can create value across the three cases and showed that leading open-weight models provide capabilities close to those of proprietary models on the relevant benchmarks. This chapter turns to the cost of accessing those capabilities and to a puzzle that emerges when price and capability are considered together.

The pattern is straightforward. For the software engineering, consumer banking and financial analysis tasks examined here, open-weight models deliver capabilities close to the proprietary frontier at a fraction of the price. This is not a marginal difference of a few percentage points; across the models compared, the difference in the price of an output token can be an order of magnitude or more. Nor is the pattern confined to the three use cases. Plotted across the broader capability spectrum, proprietary models occupy a relatively narrow high-price band that is largely independent of their capabilities, while open-weight models form a broad lower-price frontier beneath them.

If price and capability were the only variables that mattered, the market would be shifting toward open-weight models quickly. It is shifting, but slowly. Token volumes on open inference markets are rising and new routing marketplaces continue to appear, yet the revenues of frontier AI labs and their bundled distribution channels still exceed the entire volume of the open inference market by a wide margin. Capability parity at a lower price has not, so far, been sufficient to move the market rapidly toward open-weight models.

The explanation lies partly in characteristics of the service that are not captured by model benchmarks or output token prices. Frontier providers combine their models with subscriptions, established tooling and communities, tightly integrated harnesses and trusted infrastructure. They also benefit from an installed base of firms that have already built workflows around their services. Using an open-weight alternative may therefore entail switching costs and require organizational capabilities that partially offset the immediate price advantage. The following sections examine these factors in turn.

5.1 | The price of the models examined in Chapter 4

Chapter 4 showed that leading open-weight models can provide capabilities close to the proprietary frontier for the software engineering, consumer banking and financial analysis tasks examined in this report. This section compares the listed prices for accessing those models, measured in U.S. dollars per million output tokens.

Table 5 | Benchmark performance and output token prices of leading proprietary and open-weight models across the examined cases

Use case	Benchmark	Leading proprietary model and score	Leading open-weight model and score	Capability gap (pp)	Output-token price
Software production	SWE-rebench v2	Fable 5 64.5%	GLM-5.2 62.9%	1.6 pp	Fable 5: $50 GLM 5.2: $1.56
Consumer banking	τ^3-Banking	Qwen3.8 Max 51.3%	GLM 5.3 (max) 50.3%	1.0 pp	Qwen3.8 Max: $6 GLM 5.3: $3.41
Private equity	Finance Agent v2	Gemini 3.8 Flash 61.44%	GLM 5.3 Flash 57.85%	3.59 pp	Gemini 3.8 Flash: $3.75 GLM 5.3 Flash: $0.25

Note: Score gaps are expressed in percentage points (pp). Prices are in U.S. dollars per million output tokens and reflect selected provider rates listed on OpenRouter, checked on 10 September 2026.

Sources: Benchmark scores reported in Chapter 4; OpenRouter list prices.

The selected open-weight models have lower output token prices in all three comparisons, although the size of the advantage differs. In software production, GLM-5.2 performs 1.6 percentage points below Fable 5 while its output token price is approximately one-thirty-second as high. For financial analysis, GLM-5.3 Flash is 3.59 percentage points behind Gemini 3.8 Flash and is available at one-fifteenth of its output token price. The difference is smaller in consumer banking, where GLM-5.3 performs 1 percentage point below Qwen3.8 Max and has an output token price approximately 43% lower.

The price differences are particularly striking when set against the small gaps in benchmark performance. In the software production and financial analysis comparisons, proprietary output token prices are more than an order of magnitude higher. Even in consumer banking, the open-weight alternative costs substantially less. For European firms, this makes the choice of model economically consequential. Where an open-weight model meets the requirements of a workflow, the additional capability of a proprietary model must justify a substantial price premium. Access to competitively priced open-weight alternatives therefore gives firms greater scope to reduce inference expenditure and retain more of the value created by AI.

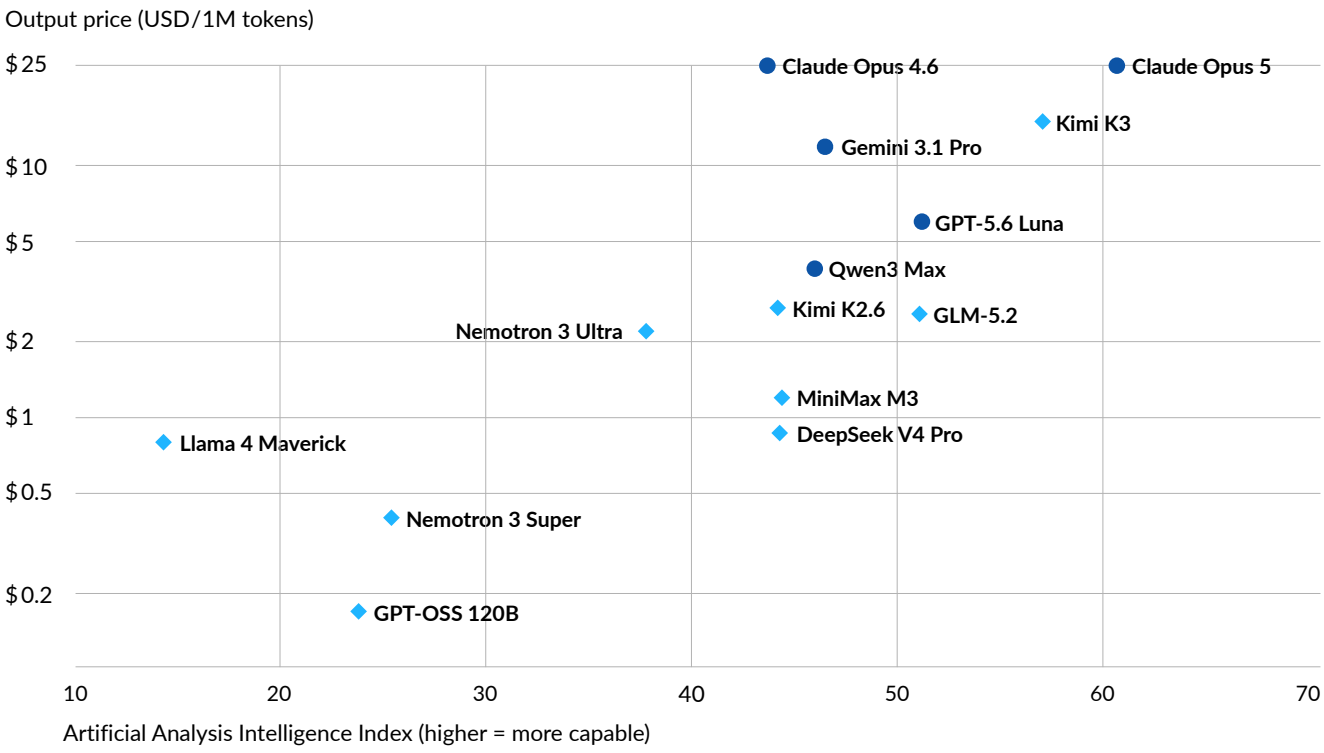
These three comparisons are based on task-specific benchmarks and therefore cover only the models examined in the respective use cases. To assess whether the same capability-price relationship extends beyond these cases, we compare a broader set of proprietary and open-weight models using a common measure of model capability.

5.2 | The broader capability-price relationship

Using the Artificial Analysis Intelligence Index as a common measure of model capability and OpenRouter prices per million output tokens, the broader comparison shows the same pattern. Proprietary models occupy a relatively narrow high-price band ($6–$25) largely independent of capability, while open-weight models form a broad lower-price frontier across the capability spectrum. Kimi K3 nearly matches Claude Opus 5's capability at about 60% of the price, while GPT-OSS 120B and Nemotron 3 Super deliver midtier capability at $0.17–$0.40.[53]

Figure 1 | Model capability and output token prices: proprietary and open-weight models

Artificial Analysis Intelligence Index against OpenRouter output-token price (USD/1M tokens, log scale)



Note: Model capability, measured by the Artificial Analysis Intelligence Index (x-axis), against OpenRouter output token list price in U.S. dollars per 1 million tokens (y-axis, log scale). Comparison of 14 models, July 2026 snapshot. More capable models tend to cost more, but the relationship is weak: Proprietary models (dark-blue circles) sit almost entirely in the $6-$25 band, while open-weight models (light-blue diamonds) span the full capability range at markedly lower prices. Several open-weight models rival mid- to upper-tier proprietary models on capability at a fraction of the price (e.g., GLM-5.2 and GPT-5.6 Luna have similar index scores, with GLM-5.2 priced at roughly one-third of the price; Kimi K3 approaches the frontier at about 60% of Claude Opus 5's price). The two highest-priced models both sit at $25 despite an approximately 17-point capability gap. Results are indicative and depend on version and date.

Sources: Artificial Analysis Intelligence Index; OpenRouter list prices.

53 More capable models generally command higher per-token prices than less capable ones, but the relationship is positive and noisy. Plotting each model's capability – measured using a benchmark such as the composite Artificial Analysis Intelligence Index – against its list price per million output tokens shows that proprietary market power, not capability alone, drives a substantial part of the price. Indicative only; results are dependent on the index version and date.

The broader comparison therefore reinforces the finding from the three use cases. High model capability is available at very different prices, and leading open-weight models occupy much of the lower-priced part of the market. Yet the higher prices of proprietary models have not eliminated demand for frontier AI services. OpenRouter shows increasing use of open-weight models but captures only part of the market because it excludes much of the direct and bundled inference sold through frontier AI labs and their cloud partners. According to The Information, OpenAI and Anthropic recorded $4.5 billion in revenue from resellers and bundled offerings, as well as $6 billion from direct API use in 2025.[54] This exceeds by a wide margin the total trading volume of OpenRouter, which we estimate at $2.54 billion.[55]

The persistence of this demand despite the price differences shown above suggests that firms are buying more than model capability at a listed token price.

5.3 | Sources of sustained competitive advantage

Before capable open-weight alternatives became widely available, the competitive advantage of frontier labs was straightforward: Their models provided the strongest available capabilities and faced limited competition. As this paper has shown, this situation has changed. By 2026, leading open-weight models had reached near-parity with proprietary models on several relevant capabilities while being available at more competitive prices. At the same time, the inference prices shown in Figure 1 suggest that proprietary frontier services continue to command substantial price premiums. Our interviews and supporting literature point to four additional features of the service offering that help explain why firms continue to use proprietary frontier AI services despite higher API prices.

Subscriptions: Our interviews with software developers (see Section 4.1) suggest that subscriptions from frontier AI labs provide a strong incentive to remain within their ecosystems. Users receive access to highly capable models and associated tools for a fixed monthly payment, making expenditure predictable and reducing the need to monitor token consumption during routine use.

While it is unclear how many tokens are included in a subscription, a simple calculation illustrates the level of discounting that may be available to intensive users of Anthropic's $200 Claude "Max 20x" plan. The software developers in our interview sample reported average monthly usage of around 30 million Opus output tokens.[56] At published API prices, this would be equivalent to approximately $750 in token costs.[57] Considering that the subscription also includes usage of smaller models such as Sonnet, the total value of the included usage may be even higher.

Subscriptions can therefore make proprietary frontier AI services considerably cheaper for individual users than their published API prices suggest. As usage expands, however, firms may eventually move beyond subscription allowances and toward API-based consumption. At that point, the substantially higher inference prices shown in Sections 5.1 and 5.2 become more important.

Subscriptions therefore help frontier AI providers acquire users before high-volume inference costs become decisive. Whether those users subsequently move to a cheaper model depends on the wider ecosystem they have already adopted and the cost of leaving it.

54 Muppidi, S. (2026), "Anthropic Lowers Gross Margin Projection as Revenue Skyrockets," *The Information*, https://www.theinformation.com/articles/anthropic-lowers-profit-margin-projection-revenue-skyrockets.

55 Based on public sources, OpenRouter's revenue is estimated at $140 million per year. According to its website, OpenRouter collects a 5.5% transaction fee, and we assume for our estimation that all revenue is derived from this fee. See Efrati, A. and Palazzolo, S. (2026), "OpenRouter Financials Suggest Steep Price For Possible Acquirer Stripe," *The Information*, 29 July. https://www.theinformation.com/articles/openrouter-financials-suggest-steep-price-possible-acquirer-stripe; OpenRouter (n.d.), "Pricing." https://openrouter.ai/pricing. Accessed 7 October 2026.

56 This figure is the average monthly usage reported by the software developers we interviewed, based on their actual Claude Code usage. Because our sample consists of experienced software developers, it likely represents relatively intensive users and should not be interpreted as average usage across Max subscribers.

57 Anthropic (n.d.), "Pricing," API section. https://claude.com/pricing#api. Accessed 7 October 2026.

Ecosystem: A GitHub search identified approximately 1.4 million repositories related to Claude Code skills, templates and other tooling. The same search for OpenCode, a coding harness for open-weight models, shows only 33,400 repositories. Similarly, Reddit communities, AI influencers and YouTube tutorials focus heavily on using Claude Code or OpenAI's Codex, indicating a large community around frontier AI services.

This ecosystem reduces the effort required to adopt and extend the services. Developers can draw on existing examples, integrations, templates and accumulated knowledge rather than building each component themselves. As Pecher et al. have shown, the strength of the supporting community can be a key factor in users' choice of services.[58]

Open harnesses such as OpenCode and Pi can provide comparable forms of functionality around open-weight models, but they do not yet benefit from a supporting ecosystem of the same scale. The resulting difference concerns the surrounding resources available to the user as much as the underlying model.

Harnesses: Harnesses connect an AI model to the tools and digital infrastructure through which it acts. A model may formulate an intent – for example, to read a file, execute code or call a tool – while the harness translates that intent into an executable action and returns the result to the model. This is known as "tool calling" or "function calling."

Within frontier ecosystems, harnesses are tailored to specific models, such as Claude Code for Claude and Codex for OpenAI's GPT models. This tight integration can improve efficiency and quality. Both use provider-specific approaches that are closely integrated with the underlying models, creating a potential source of competitive advantage.[59]

Open harnesses also support tool calling for open-weight models, but they need to accommodate differences among models because the way models express tool-use intentions is neither standardized nor harmonized. Benchmarks focused specifically on tool-calling capabilities, such as Berkeley's BFCL, show that there is still a real gap in reliable tool use between frontier and open-weight models – accuracy can vary by up to 20%.[60] Improved harnesses and stricter configurations can offset part of this difference, but doing so requires additional engineering by the user. Frontier providers therefore compete through the integration of model and harness as well as through the capability of the model itself.

Trust: The final service-level consideration is trust in the inference provider. Pecher et al. identify privacy as an important consideration in decisions to switch services.[61] Models themselves pose limited direct privacy risk because they do not independently record user interactions or messages. The inference providers serving them, however, process customer inputs under provider-specific retention and security arrangements. The relevant question therefore becomes whether users trust the inference provider.

Findings from our interviews indicate that this consideration is particularly important in financial services. Interviewees expressed greater confidence in accessing frontier AI models through established cloud infrastructure such as Microsoft Azure, Amazon Bedrock or Google Vertex than through less familiar inference providers. Established contractual relationships, security processes and compliance arrangements therefore form part of the service being purchased.

Within Europe, questions about trust in U.S.-based cloud providers are also central to the current digital sovereignty debate.[62]

58 Pecher, J., Syrmoudis, E. and Grossklags, J. (2024), "Service selection and switching decisions: user behavior in high-interoperability environments," *Humanities and Social Sciences Communications*, 11, article 1585. https://doi.org/10.1057/s41599-024-04056-4.

59 Anthropic's Programmatic Tool Calling and OpenAI's Function Calling are well documented. See Anthropic (n.d.), "Programmatic tool calling." https://platform.claude.com/docs/en/agents-and-tools/tool-use/programmatic-tool-calling. Accessed 7 October 2026; OpenAI (n.d.), "Function calling." https://developers.openai.com/api/docs/guides/function-calling?api-mode=responses. Accessed 10 September 2026.

60 University of California, Berkeley (2026), "Berkeley Function Calling Leaderboard (BFCL) V4." https://gorilla.cs.berkeley.edu/leaderboard.html. Accessed 10 September 2026.

61 Pecher, J., Syrmoudis, E. and Grossklags, J. (2024), "Service selection and switching decisions: user behavior in high-interoperability environments," *Humanities and Social Sciences Communications*, 11, article 1585. https://doi.org/10.1057/s41599-024-04056-4.

62 Hessel, S. (2025), "Digitale Souveränität: Wie souverän sind die EU-Angebote der Hyperscaler?" *heise online*, 27 June, https://www.heise.de/hintergrund/Digitale-Souveraenitaet-Wie-souveraen-sind-die-EU-Angebote-der-Hyperscaler-10461972.html.

Open inference markets therefore face a dual challenge. They need to compete on price while also providing the legal, security and compliance assurances required by enterprise customers. OpenRouter has begun to provide information about provider data retention practices, but some inference providers still provide limited information on their legal entity, privacy policies, security practices or certifications.[63] This can constrain enterprise use of otherwise attractive open-weight models, particularly in regulated sectors such as finance, insurance and energy.

5.4 | Why lower prices have not led to faster adoption

The service-level advantages described above help explain why lower inference prices do not translate immediately into a shift toward open-weight models. Two additional barriers arise on the demand side. Firms may already have substantial investments in frontier AI ecosystems, and obtaining the full economic benefit of open-weight models may require capabilities that many organizations have not yet developed.

First-mover advantage: Capability parity between leading open-weight and proprietary models is a relatively recent development, gaining traction with the release of models such as DeepSeek, GLM-5.2, MiniMax, Kimi K2 and Qwen3. Frontier AI providers had several years in which they offered the strongest available models and firms built applications, agents and workflows around their services. This installed base gives them an important first-mover advantage. Moving to an open-weight alternative can require firms to change integrations, evaluate the new model, adapt prompts and harnesses, repeat security and compliance checks and retrain employees. Pecher et al. find that the time required to switch can materially affect users' willingness to change digital services.[64] The relevant comparison for an adopting firm therefore includes both the expected savings from lower inference prices and the time and resources required to realize them.

Organizational capabilities: The second barrier concerns the capabilities of the adopting organization. Open-weight models create additional options because firms can combine them with their own data, infrastructure and workflow-specific engineering. Revolut's PRAGMA model illustrates the potential value of this approach. As discussed in Chapter 4, Revolut built a specialized model around proprietary banking event data and combined it with open-weight components to improve performance on economically important banking tasks.

Realizing this type of advantage requires more internal capability than purchasing access to a ready-made frontier AI service. Firms need expertise in model selection, evaluation, data engineering, deployment and, for more specialized applications, model development. Where those capabilities are absent, the convenience of an integrated proprietary service can outweigh lower inference prices.

This distinction helps explain the central finding of this chapter. Open-weight models already provide strong capabilities at substantially lower inference prices, but the price of the model is only one part of the adoption decision. Subscriptions, surrounding ecosystems, integrated harnesses and trusted infrastructure increase the value of proprietary services, while switching costs and limited organizational capabilities slow the move toward open-weight alternatives.

For Europe, the implication is that access to capable and competitively priced open-weight models is a necessary starting point, but firms must also be able to use them effectively.

63 Moberg, C. (2026), "Guardrails: Protect your Agents, Data, and Costs," *OpenRouter*, 29 May (updated 24 September), https://openrouter.ai/blog/announcements/guardrails/.

64 Pecher, J., Syrmoudis, E. and Grossklags, J. (2024), "Service selection and switching decisions: user behavior in high-interoperability environments," *Humanities and Social Sciences Communications*, 11, article 1585. https://doi.org/10.1057/s41599-024-04056-4.

6 | Discussion

Across the use cases examined in this report, leading open-weight models already provide capabilities close to those of proprietary frontier models. At the same time, they can be accessed at substantially lower inference prices. Lower inference prices can allow European firms to retain more of the value their applications create. Chapter 5 shows that realizing this advantage also depends on the services through which models are accessed and the capabilities of the adopting organization. Europe's AI strategy should therefore strengthen the tools, trusted services and deployment expertise that allow firms to use open-weight models effectively.

This matters because Europe currently trails the leading developers of frontier models by a substantial margin.[65] Europe's economic position will therefore depend not only on frontier model development but also on whether European firms can access strong AI capabilities on competitive terms and use them to improve their own products and operations. If capable models are available at lower prices, more of the productivity gains from AI can accrue to European firms and their customers rather than to a small number of vertically integrated model providers. Moreover, if European firms develop the capabilities to tailor open-weight models to specific economic tasks, they may be able to further reduce operating costs, improve quality and develop sources of competitive advantage.

Open-weight models create opportunities for European providers to supply these services. Where licenses permit independent hosting, European cloud providers and specialized inference providers can serve the same underlying model and compete on price, speed, reliability and trust. Systems integrators and application developers can build sector-specific services around these models without first financing frontier model development. Value creation can therefore take place not only at the model layer but also in inference, deployment, adaptation and integration into business processes.

For Europe, a relevant strategic objective is to maintain a capable and competitive open-model ecosystem. The nationality of an individual model developer is secondary to whether European firms can host and adapt models and switch between models and providers. A capable open-weight model developed outside Europe may strengthen the position of European firms more than a proprietary European model that can be accessed only through one vertically integrated provider. At the same time, Europe cannot determine whether developers outside Europe will continue to release capable models with open weights. Maintaining a capable European open-model ecosystem can therefore provide resilience against future restrictions on the availability or commercial use of foreign open-weight models.

This also strengthens Europe's control over AI deployment. Open models give firms and public institutions greater scope to evaluate models independently, adapt them to European languages, sectoral requirements and value cases and deploy them on their own infrastructure or through European providers. They can therefore support both the competition and accountability objectives set out in Chapter 2. These benefits are particularly relevant in Europe, where firms operate across multiple languages and regulated sectors and where control over data and infrastructure has become an important policy objective.

65 European Commission, Directorate-General for Communications Networks, Content and Technology (2026), *Enhancing Competitiveness, Sovereignty and Security of the European Union in Frontier AI: Key Insights from the First Meeting of the European Expert Forum on Frontier AI*, Luxembourg: Publications Office of the European Union. https://doi.org/10.2759/8947951.

This also changes the role of public compute and AI infrastructure policy. Public support should not be directed only toward training ever-larger frontier models. It should also support fine-tuning, evaluation and inference for open models and enable European providers to compete in serving them. Access to public compute should be widespread, serving European firms of all sizes, and should be coupled with access to the technical capabilities required to tailor open-weight models to firm- or sector-specific value cases. The objective is to ensure that access to capable AI inference does not become tied to a small number of proprietary model and cloud ecosystems and that firms across the European economy develop the capabilities needed to apply AI productively.

The economic case is therefore broader than technological sovereignty understood as European ownership of frontier models. Europe benefits when its firms have access to capable models at competitive prices, when several providers can supply those models and when firms can move between them without rebuilding their AI applications. Open-weight models, in conjunction with open, competitive and trusted markets, can provide that alternative. The next chapter examines the risks to the licenses, inference markets and financing arrangements on which this opportunity depends.

7 | Risks to the current open-model price advantage

Chapters 4 and 5 identify a current advantage for open-weight models. Leading open-weight models provide capabilities close to those of proprietary frontier models for the sectors examined while being available at lower inference prices. This allows European firms to retain more of the value they create with AI and gives European providers opportunities to compete in inference, adaptation and deployment.

This advantage rests on a particular market configuration. Model licenses must allow independent providers to sell inference. Several providers must be able to serve the same models and compete on price. Model development and inference provision must also generate sufficient revenue to remain viable. Each of these conditions may change even if capable open-weight models remain available.

7.1 | The advantage depends on licenses that permit competitive inference

The publication of model weights creates the potential for inference competition where the license permits independent providers to host the model and sell access to it. Where this right is available without provider-specific approval, several companies can serve the same model and compete on price and service quality. Where providing commercial inference requires a separate agreement, the model developer retains greater control over entry into the inference market.

Several current licenses provide broad commercial rights. Nvidia's Open Model Agreement grants no-charge and royalty-free rights to reproduce, sublicense and distribute the covered model materials. OpenMDW-1.1 similarly permits the model materials released under it to be used and distributed without restriction. These licenses allow an inference provider to serve a model without first negotiating a bilateral agreement with its developer.[66]

Recent releases show that these terms can change between model generations. Kimi K2 was released under a modified MIT license that broadly permitted commercial use and distribution. Its additional condition concerned attribution by very large products and services. The Kimi K3 license retains broad rights over the weights but introduces a separate requirement for companies operating a model-as-a-service business. Once the company and its affiliates exceed $20 million in aggregate revenue over a consecutive 12-month period, they must conclude an additional agreement with Moonshot AI before using Kimi K3 commercially.[67]

This represents a significant shift in the structure of commercial inference. Kimi K3 remains an open-weight model, but large independent providers no longer have an unconditional right to serve it. By allowing Moonshot AI greater control over which

66 NVIDIA's Nemotron Open Model License grants a perpetual, worldwide, no-charge and royalty-free license to reproduce, modify, sublicense and distribute covered models and derivative works and explicitly permits commercial use. OpenMDW-1.1 similarly grants unrestricted, royalty-free permission to use, modify and distribute the model materials released under it. See NVIDIA (2025), "NVIDIA Nemotron Open Model License"; OpenMDW (n.d.), "OpenMDW-1.1."

67 Moonshot AI's Kimi K2 was released under a modified MIT license that broadly permits commercial use and distribution; its principal additional condition requires attribution for products or services exceeding 100 million monthly active users or $20 million in monthly revenue. The Kimi K3 license introduces an additional restriction for model-as-a-service providers. Where the licensee and its affiliates operate a model-as-a-service business and exceed $20 million in aggregate revenue over any consecutive 12-month period, they must enter into a separate agreement with Moonshot AI before using Kimi K3 commercially. Internal use and use through Moonshot AI's official products or certified inference partners are exempt. See Moonshot AI, Kimi K2 License; Moonshot AI, Kimi K3 License.

inference providers can offer the model, the license could limit competition and potentially increase inference prices.

Similar license terms would weaken the price mechanism identified in Chapter 5. Separate agreements, license fees or partner certification could reduce the number of providers serving the most capable open-weight models, increasing market concentration.

Any additional costs may also be reflected in the price paid by users. The open-weight label would remain, while the market for serving the model would begin to resemble the controlled distribution system of frontier AI services.[68]

The current price advantage therefore depends on future capable models continuing to be released under licenses that permit independent commercial inference.

7.2 | The advantage depends on an inference market with multiple providers

Permissive licenses create the legal basis for competition. The price effect requires several providers with the infrastructure and technical capacity to serve the same model. Firms must also be able to change providers without changing the underlying model or rebuilding the application.

The current market shows that this structure can work. OpenRouter lists the same open-weight models through multiple providers and makes their prices, throughput and availability directly comparable. GLM-5.2, for example, is currently offered by 24 providers at different prices. Users can therefore select an inference provider independently of their choice of model.[69]

European companies are also beginning to establish themselves in this part of the value chain. In August 2026, Mistral expanded its inference platform to third-party open models, beginning with Z.ai's GLM-5.2. Mistral serves the model without modification and provides access through the same infrastructure, regional controls and service arrangements used for its own models. This allows Mistral to compete by providing trusted infrastructure and services for a model developed elsewhere.[70]

This development is important for Europe. A European provider does not need to finance the development of every model it offers. It can create value by giving European customers reliable access to capable models and by competing on the conditions under which those models are served. The separation between model development and inference provision therefore opens part of the AI value chain to European firms even when the underlying model was developed outside Europe.

Multiple distribution channels do not always produce this type of competition. Anthropic's Claude models are available directly and through AWS and Microsoft Foundry. In both partner environments, however,

68 Anthropic's distribution through several cloud platforms does not create independent competition over its underlying per-model token rates. Microsoft states that Claude usage in Foundry is converted into Claude Consumption Units using Anthropic's published per-model token rates, after any contractual discounts, and that the same CCU construct is used across Anthropic's supported cloud platforms. AWS similarly prices token usage at Anthropic's standard per-model rates before applying negotiated discounts and converting the resulting amount into CCUs. See Microsoft (2026), "Claude consumption units (CCU) billing in Microsoft Foundry," last updated 15 August. https://learn.microsoft.com/en-us/azure/foundry/foundry-models/concepts/claude-models-billing. Accessed 7 October 2026; Amazon Web Services (n.d.), "Billing," *Claude Platform on AWS User Guide.* https://docs.aws.amazon.com/claude-platform/latest/userguide/billing.html. Accessed 7 October 2026.

69 OpenRouter illustrates competition between inference providers serving the same open-weight model. For GLM 5.2, several providers offer the same model at different listed prices and with different latency, throughput and uptime. See OpenRouter (n.d.), "GLM 5.2 – API Pricing & Benchmarks," provider comparison. https://openrouter.ai/z-ai/glm-5.2. Accessed 1 October 2026.

70 Mistral announced in August 2026 that its platform would begin supporting third-party open models, starting with Z.ai's GLM 5.2. Mistral states that GLM 5.2 is hosted without modification and runs through the same infrastructure, regional controls and service arrangements as its own models. See Mistral AI (2026), "In-region inference, open models, and new European infrastructure for sovereign AI," 11 August. https://mistral.ai/news/regional-inference-open-models-new-compute/; Mistral AI (n.d.), "Z.ai GLM 5.2." https://docs.mistral.ai/models/zai-glm-5-2. Accessed 1 October 2026.

token use is priced using Anthropic's published per-model rates before being converted into Claude Consumption Units. Contractual discounts may differ, but the cloud providers do not independently determine the underlying price per token. The additional distribution platforms expand access to Claude without creating an open market in which any provider can serve the model and compete on price.[71]

The open-weight inference market may nevertheless become more concentrated. Serving large models requires substantial computing capacity, specialized software and sufficient demand to operate infrastructure efficiently. These requirements favor providers with scale and financial backers willing to invest large amounts of capital upfront. Model developers may also direct demand toward first-party services or certified partners, while provider-specific interfaces, security systems and development tools can make switching more difficult.

A market with several nominal providers can therefore provide limited competitive pressure if customers remain tied to a single ecosystem, harness and set of services. An integrated provider may also retain customers after raising prices where it combines model access with cloud infrastructure and development tools, such as harnesses or agent development kits. The current advantage depends on continued entry by inference providers and on firms' practical ability to move between them. The latter requires sufficient harmonization and standardization to ensure interoperability.[72]

7.3 | Current prices may not be durable

The final risk concerns the economics of model development and inference provision. Open-weight models lower inference prices for adopting firms, but independent inference revenue does not automatically accrue to the developer of the underlying model. Developers must still finance model development, particularly the costly stages of pretraining, midtraining and post-training.

Several financing arrangements currently coexist. First, open-model developers can sell inference directly through their own APIs, provide enterprise services or offer complementary products. This enables them to capture more value and potentially earn higher margins that can be reinvested in capability development. DeepSeek and Mistral, for example, combine new model releases with paid first-party inference on their own infrastructure. Second, public authorities can subsidize model development and computing capacity. Third, hardware companies have an incentive to support open models because wider adoption increases demand for chips, inference software and related infrastructure. Nvidia's Nemotron program follows this model by combining open model releases with its wider hardware and deployment business.[73]

71 See footnote 68.

72 Baur, H., Pompéry, M. and Sieker, F. (2025), *The Standard of Standards: Digital Sovereignty and Fair Competition Through Cross-Sector Interoperability*, Gütersloh: Bertelsmann Stiftung. https://doi.org/10.11586/2025110.

73 Open-model development is currently financed through several channels. Developers can combine open-weight releases with paid first-party inference; DeepSeek and Mistral, for example, operate APIs with published per-token prices. Public authorities can subsidize model development and computing capacity, including through Chinese municipal "computing-power vouchers" and direct compute subsidies. Hardware companies also have an incentive to support open models because wider model adoption increases demand for inference hardware and software; NVIDIA's Nemotron program is a prominent example. See DeepSeek (n.d.), "Models & Pricing." https://api-docs.deepseek.com/quick_start/pricing/. Accessed 1 October 2026; Mistral AI (n.d.), "Pricing." https://docs.mistral.ai/inference/pricing. Accessed 1 October 2026; Grimm, S. (2025), "China subsidizes AI computing for small domestic companies — 'computing power vouchers' spread across multiple Chinese cities," *Tom's Hardware*, 3 September. https://www.tomshardware.com/tech-industry/artificial-intelligence/china-subsidizes-ai-computing-for-small-domestic-companies-computing-power-vouchers-spread-across-multiple-chinese-cities.

Model developers seeking to capture more value have several options. First, they can charge inference providers for commercial access through license fees, require separate agreements above a specified scale or direct users toward certified partners. Second, they can reserve their most capable models exclusively for their own APIs and infrastructure or delay the release of open-weight versions. The Kimi K3 license provides an early example of a developer retaining open weights while increasing its control over commercial inference by third parties.

Should bundling become more widespread, it would alter the present division of value. A developer that controls both the model and its inference infrastructure can use inference revenue to finance further model development. At the same time, vertical integration reduces competition among inference providers and allows the integrated company to capture a larger share of the value generated by model access. This would move the market closer to the proprietary frontier-provider model examined in Chapter 2.

Inference providers face a related problem. Competitive prices benefit users, but providers must cover the cost of computing capacity, energy, software, reliability and customer support. Providers unable to operate profitably may leave the market or be acquired. Greater concentration would weaken price competition even where model licenses remain permissive.

The stability of the current configuration therefore depends on several unresolved questions: who ultimately pays for model development, how far current prices reflect subsidies or strategic pricing and how revenue is divided between model developers and inference providers. It will also depend on whether developers change their licenses, move toward bundling or reserve their most capable models for controlled services. At the inference layer, the central question is whether several independent providers can remain profitable and whether an integrated provider could raise prices while remaining competitive.

8 | Conclusion and outlook

This paper has examined how firms create economic value with AI and which model capabilities they require. Across software production, consumer banking and private equity, the clearest evidence concerns cost advantage through higher productivity. Software developers and investment professionals report time savings, while automated customer service offers banks a route to lower operating costs. The cases also identify potential for differentiation through better products, services and investment decisions but provide little evidence that this potential has yet translated into sustained commercial benefits.

The benchmark and price comparisons identify leading open-weight models that provide strong capabilities relevant to these applications at substantially lower inference prices. This gives European firms an opportunity to reduce expenditure on model access and retain more of the value they create.

These benefits depend on the market conditions that currently make capable open-weight models available at low prices and through competing providers. Chapter 7 therefore examines three risks to this configuration. Tighter licenses, fewer competing inference providers and insufficient funding for future models could weaken the combination of strong capabilities, low prices and provider choice. For Europe, the unresolved issue is whether the market configuration most favorable to adopting firms can also sustain the model development and inference provision on which that advantage depends.

The second paper examines this question by tracing where $100 of customer expenditure on AI inference goes under different market configurations. It analyzes how this expenditure is distributed among model developers, inference providers and upstream suppliers, which costs it covers and where prices fall short of the full economic cost of provision. It also examines who finances any resulting gap and how licensing arrangements, subsidies and vertical integration change the distribution of revenues and costs. Thus, while the first paper identifies the opportunity for European firms to create and retain value with AI, the second examines the economic conditions under which that opportunity can endure.

Acknowledgments

We would like to thank Maximilian Gahntz, Thieß Petersen, Charlotte Freihse, Asena Soydas, Alek Tarkowski and Christiaan van Veen for their constructive feedback and valuable contributions throughout the analysis and drafting of this study. We also thank Barbara Serfozo for her excellent editorial work, Ines Meyer and Paul Feldkamp for the design of the graphics and layout, and Franziska Damerow and Carina Wegener for their support in disseminating this publication.

Address | Contact

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telephone +49 5241 81-0
www.bertelsmann-stiftung.de

Dr. Felix Sieker
Senior Expert
Digitalization and the Common Good
Telephone: +49 30 275788-156
felix.sieker@bertelsmann-stiftung.de
www.digitalisierung-und-gemeinwohl.de/en

BertelsmannStiftung