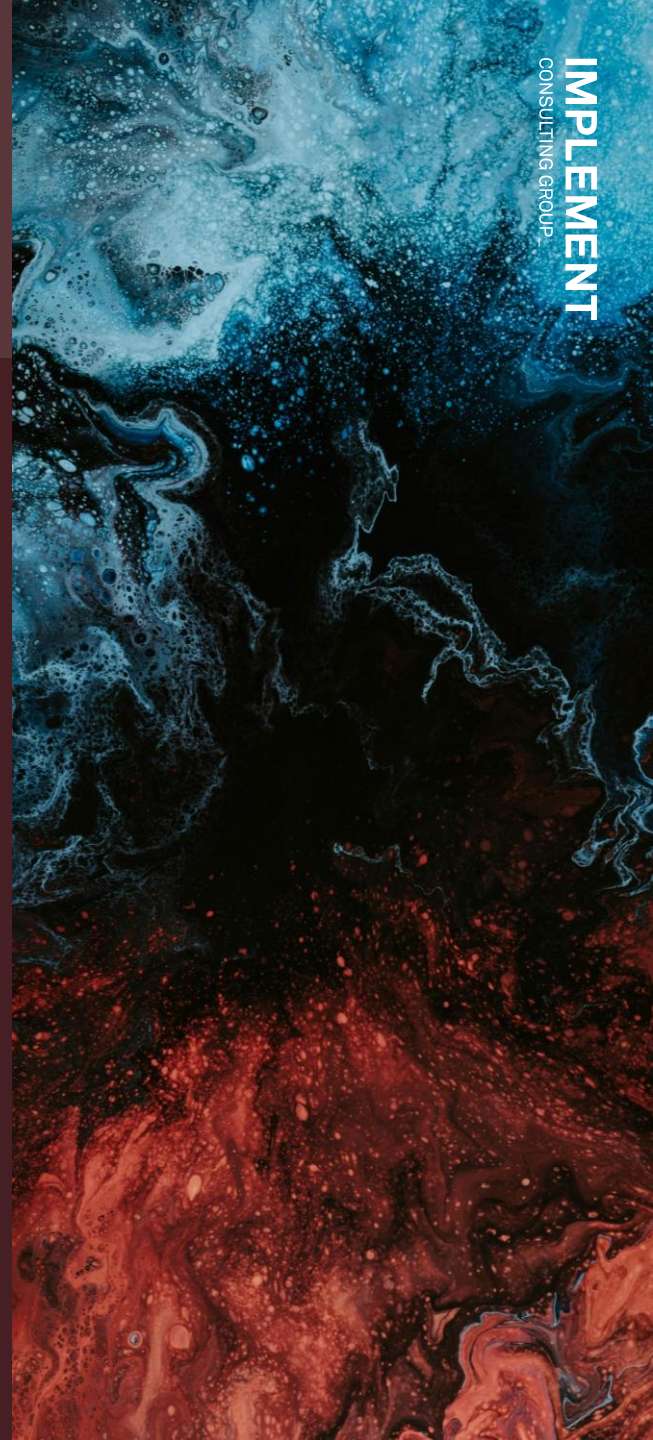
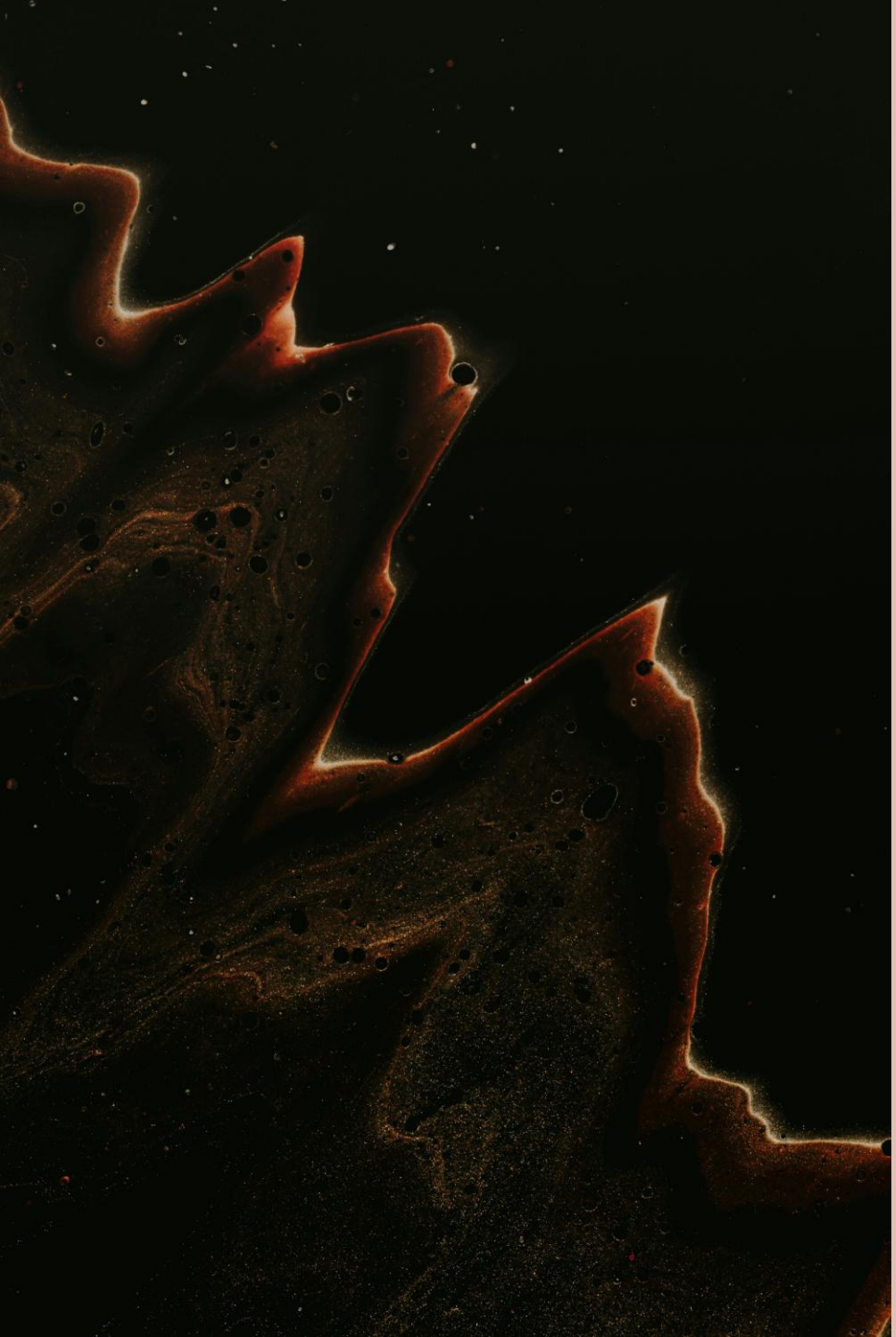
The TDM equation

Why a EUR 600 billion share of Europe's AI opportunity hinges on text and data mining (TDM)

An Implement Consulting Group study
In collaboration with Prof. Dr. Oliver Falck, Ifo Institute in Munich
Commissioned by CCIA Europe and Google

June 2026





Preface

This report assesses the economic value of text and data mining for Europe's competitiveness.

- The report focuses on commercial text and data mining (Article 4 of the CDSM Directive). Text and data mining for scientific research (Article 3) is not addressed in this report.
- Commercial text and data mining is vital for AI development, as model and application performance relies heavily on large, diverse datasets.
- This report builds on Implement's economic estimates for generative AI in the EU, estimating EUR 1.2 trillion in annual GDP gains from adopting GenAI alongside EUR 450 billion in added innovation potential.
- The report assesses the pros and cons of changes to the current commercial TDM regime, including changes to the opt-out mechanism, licensing for AI training, and transparency requirements.
- Finally, the report quantifies the value of the commercial TDM regime to Europe's competitiveness in the AI era and provides recommendations for how the commercial TDM regime can best support it.

Contributors

The authors are especially thankful to Prof. Dr. Oliver Falck from the Ifo Institute for Economic Research in Munich for valuable input and collaboration in developing the report's perspective on AI, innovation, and Europe's competitiveness. The report has also benefited from discussions on the legal regime and policy debate surrounding text and data mining for AI training with Brinkhof. The authors furthermore benefitted from discussions with Google and CCIA Europe experts during the development of the report.

Authors

This report has been prepared by the Implement Consulting Group's economics team and commissioned by CCIA Europe and Google. The authors of the report are Martin H. Thelle, Nikolaj Tranholm-Mikkelsen and Mads Sigurd Franch Andersen.

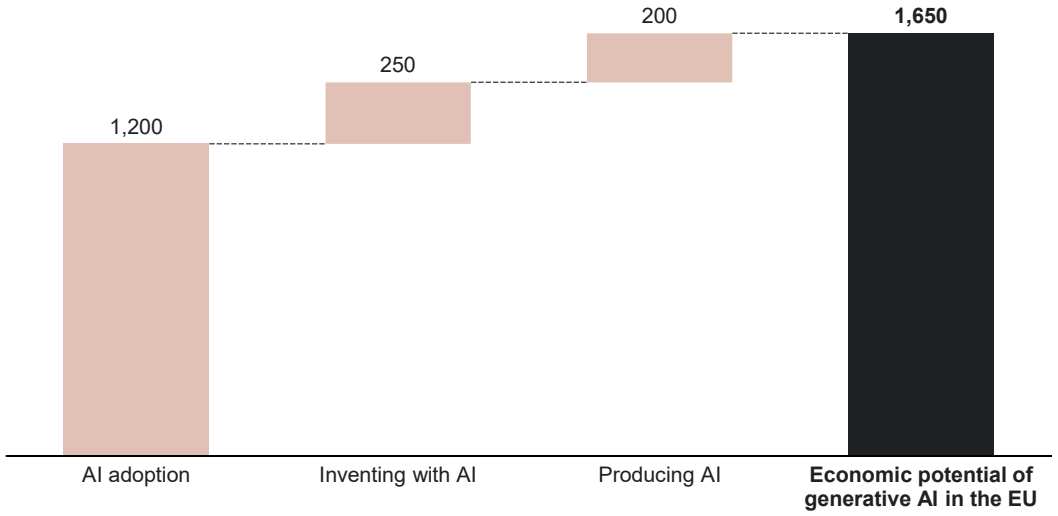
Executive summary | To fuel the AI opportunity, Europe needs a common data space with commercial text and data mining

The EUR 1.65 trillion economic opportunity…

- Adoption of generative AI offers Europe a large economic opportunity, with annual GDP gains of EUR 1.2 trillion at widespread adoption, equivalent to around 8% of GDP after a ten-year adoption period, under the current commercial TDM regime.
- Beyond adoption gains, generative AI could unlock a further EUR 200 billion from producing AI in the EU and EUR 250 billion from using AI to accelerate European research and innovation.
- Realising Europe's AI opportunity depends heavily on capabilities of frontier AI models and on the ability for European innovators to develop and fine-tune AI applications.

Economic potential of generative AI in the EU

EUR billion annually

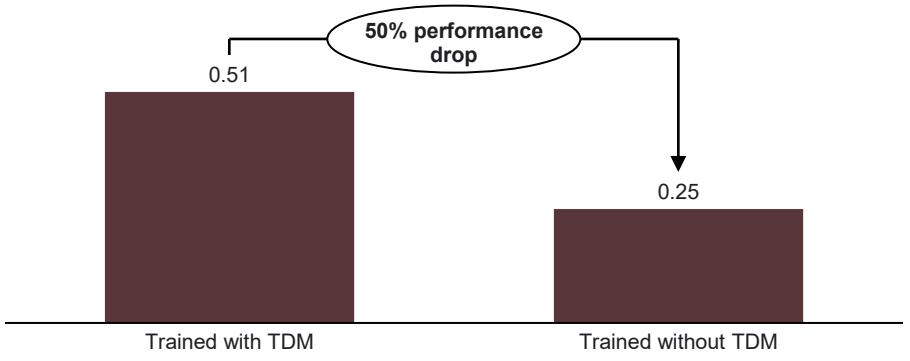


… needs a common data space with commercial TDM

- Access to a common European data space with commercial text and data mining is vital for modern AI development because frontier models and European AI applications are trained on vast datasets, and larger training datasets improve performance.
- Frontier AI models and applications are trained and fine-tuned on large datasets, without the models retaining the data themselves, and even synthetic data requires high-quality "human-mined" data.
- Training AI models on smaller and less diverse datasets weakens model capabilities materially, with benchmark tests showing a 50% drop in complex reasoning performance.
- A common European data space with large and diverse training datasets are particularly important for generative AI capabilities in smaller European languages.
- Because model capability and the ability to produce European AI application shape AI adoption and innovation, Europe needs a common data space with a balanced regime that preserves commercial TDM, practical opt-outs, and proportionate transparency requirements. This is also needed to keep AI innovation open to innovators without access to 1st party training data.

Model performance on complex reasoning

Benchmark test score (max = 1)

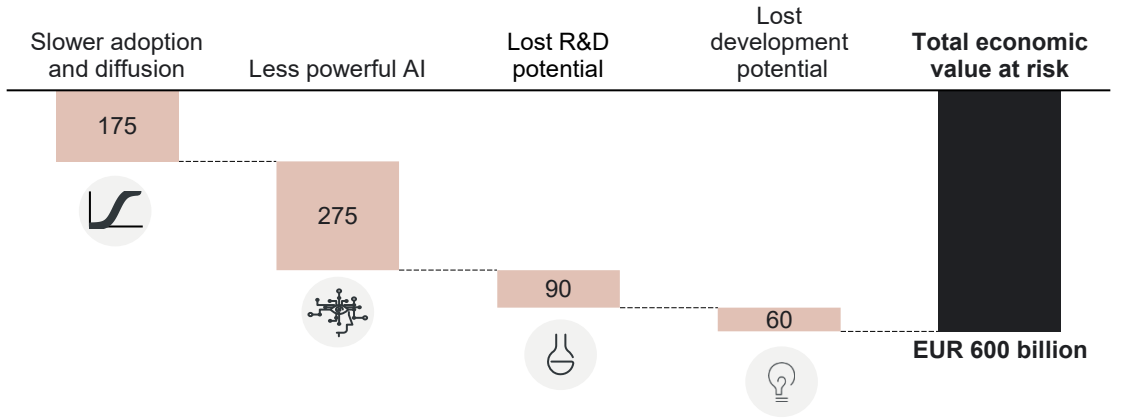


Executive summary | EUR 600 billion of European competitiveness gains are at risk from restricting the current commercial TDM regime

Competitiveness gains at risk

- The current commercial TDM exceptions were put in place to give European businesses the right conditions for innovating and staying competitive, while at the same time protecting rightsholders and incentivising new content creation. With the “fair use” regime in the US, innovators have benefitted from TDM exceptions globally. Given the rise of generative AI, the TDM regime is back under the spotlight.
- Restricting the use of commercial TDM in the EU for AI training puts EUR 600 billion of Europe’s AI opportunity at risk through slower adoption, less powerful models, lower R&D productivity, and reduced prospects for European AI development.
  - Regulatory certainty around the commercial TDM regime is needed to ensure speedy AI adoption, deployment and development of AI across Europe. Regulatory certainty on TDM is key for European AI developers and adopters to move forward with advanced AI solutions. A one-year delay in AI adoption would cost EUR 175 billion in economic value.
  - Other initiatives, such as a mandatory licensing regime, disproportionate transparency requirements and unworkable opt-out mechanisms, can lead to less powerful AI and lost R&D and development potential in the EU, putting a further EUR 425 billion at risk.
- The value at risk would fall most heavily on the EU’s high value added and knowledge-intensive industries such as pharma, professional services, and advanced manufacturing.

Economic value at risk
EUR bn per year



The way forward

- A balanced TDM regime is crucial to support European innovation and competitiveness while sustaining incentives to create new content.
- That means preserving a workable commercial TDM regime, including transparency measures that are proportionate and operationally feasible, and ensuring opt-outs are practical and machine-readable.
- Europe should also leave room for voluntary, commercial partnerships and data-sharing collaborations to develop where they add value, rather than replacing TDM with prescriptive or premature licensing interventions.

To support European innovation and competitiveness in the AI era, EU policymakers should preserve three core elements of a workable regime for commercial TDM:

Commercial TDM exception	**Maintain the current commercial TDM exception** The commercial TDM regime provides the legal certainty needed for AI development and adoption in Europe. Policymakers should preserve that baseline while ensuring that any transparency requirements are proportionate, workable, and compatible with AI development at scale.
Opt-out possibility	**Keep opt-out mechanisms workable** AI training already takes place in the EU under the current TDM Article 4 opt-out regime. Regulators should endorse the widely used existing opt-out protocol, robots.txt, since it is a scalable, machine-readable, and globally recognised standard for web crawling for AI development. Regulators should avoid immature or unproven alternative mechanisms.
Voluntary licensing	**Give room for commercial partnerships to develop further** Mandatory licensing goes against copyright frameworks that balance rightsholder protection with innovation and societal benefit. The opt-out gives rightsholders control and allows for voluntary licensing. Policymakers should continue to allow commercial partnerships and data-sharing to grow, rather than imposing mandatory or centralised models.

Contents

The economic potential of AI

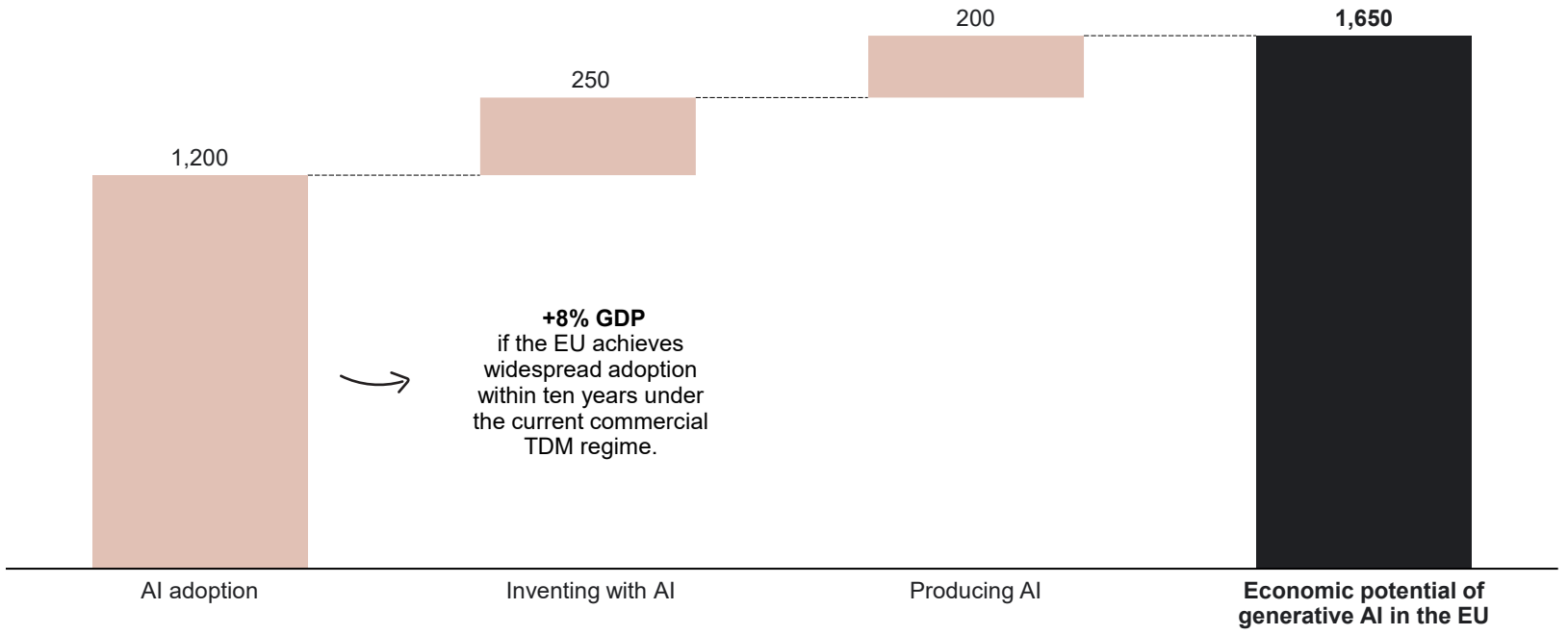
Generative AI offers the EU a large but uncertain economic potential, which depends on the pace of adoption and capabilities of models.

IMPLEMENT
CONSULTING GROUP

Generative AI offers the EU an annual EUR 1.2 trillion productivity potential and a further EUR 450 billion from AI innovation – a total of EUR 1.65 trillion

The potential impact of generative AI on GDP in the EU

EUR billion increase from baseline GDP after a ten-year adoption period



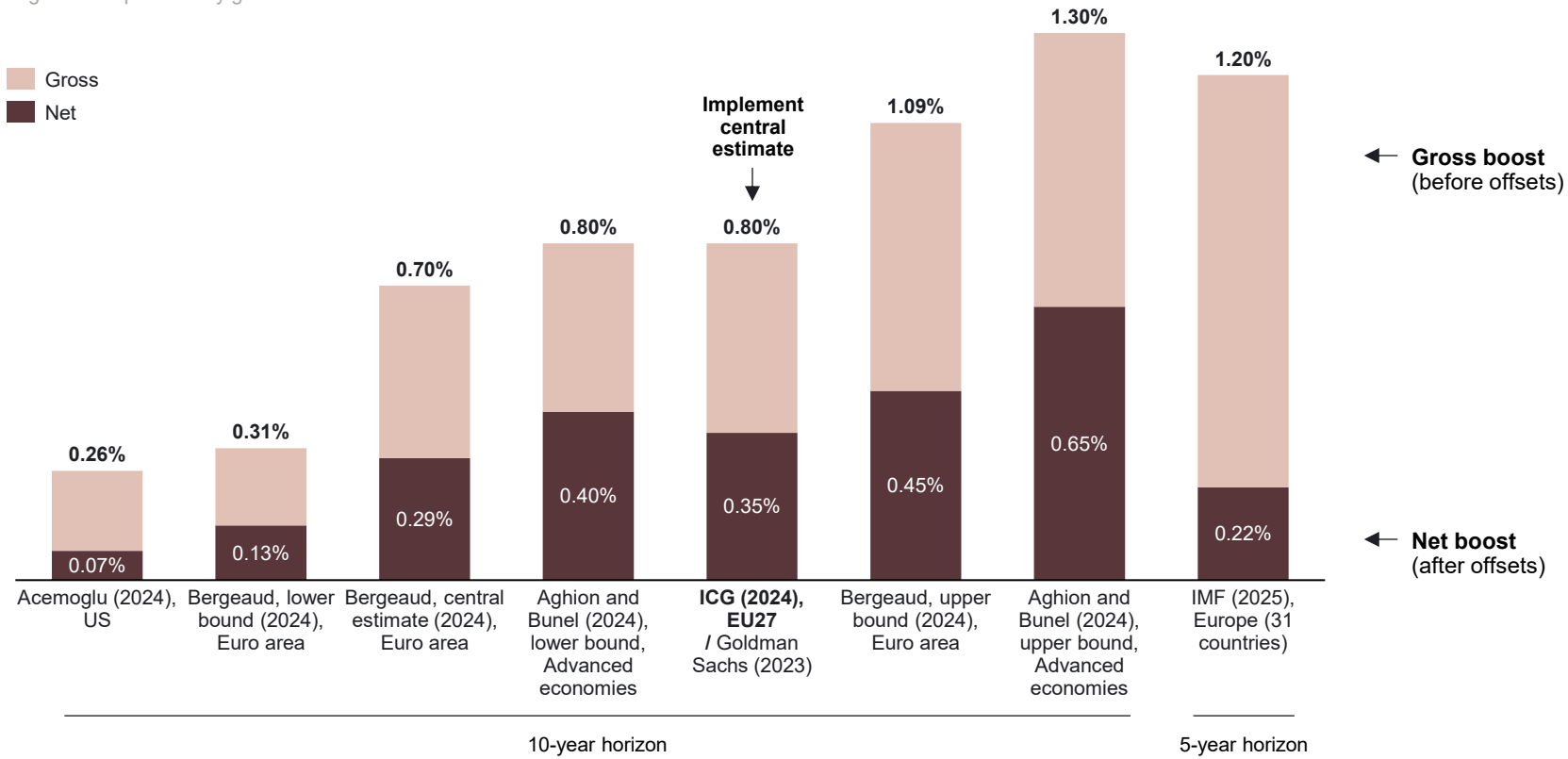
- Widespread adoption of generative AI presents a massive economic opportunity, potentially adding EUR 1.2 trillion to the EU's GDP within ten years. This growth is mainly driven by augmenting the everyday capabilities of most workers to increase overall quality and efficiency. This core estimate factors in the extra value created when AI frees up workers to focus on completely new tasks.
- Using AI to significantly accelerate research and development contributes to that innovation number. By making research much more efficient in key sectors, the EU can capture an annual GDP boost of EUR 250 billion.
- The final EUR 200 billion comes directly from the economic potential of producing AI technology. This captures the value of building the required infrastructure, core models, and specialised software services locally to secure a strong share of the global market.

Note: The estimate assumes widespread adoption of generative AI over a ten-year period. There is much uncertainty around the capability and adoption timeline of generative AI. The size of the productivity boost depends on the difficulty level of tasks that generative AI will be able to complete and the number of jobs it can automate. The estimated boost from generative AI may not be fully additive to GDP trends, as the GDP forecast already assumes a growth contribution from new technologies and generative AI may substitute some of that. Also, the boost from generative AI may be partially offset by an underlying growth slowdown. The AI production potential is based on EU's estimated market share of the forecasted global AI revenue by Bloomberg and FTI Delta. The forecasted global AI revenue excludes revenues from devices, digital ads and gaming. Foundation models relate only to Generative AI. GenAI estimates from Bloomberg are extended from 2032 to 2034 while non-GenAI estimates are extended from 2030 to 2034. EU's market shares are estimated based on various proxies for infrastructure, foundation models and apps & services. The inventing with AI potential is based on firm-level productivity gains, which are estimated on the top 800 companies in Europe (in terms of R&D expenditure) using company profits and labour remuneration from the sectoral averages through Eurostat. The productivity gains are then applied to relevant sector aggregates to reflect the entire EU economy. Sectors are drawn from McKinsey (2023) and Babina et al. (2021). For further detail on the economic potential of producing and inventing with AI, see Implement (2025).
Source: Implement Economics based on Eurostat, O*NET and Briggs and Kodnani (2023).

Studies predict AI will boost productivity, but by how much is uncertain

Implement Consulting Group, ICG (2024)
Based on the method in Briggs and Kodnani (2023), ICG (2024) analysed the automatability of tasks and productivity benefits from AI and concluded that generative AI could **raise EU GDP by 8%** over a ten-year period, with a productivity boost of **1.4 percentage points** in peak year at widespread adoption. This corresponds to an average productivity boost of 0.8% per year over the ten-year period before offsets. In line with Briggs and Kodnani (2023), ICG (2023) assessed that GenAI will to some degree offset the growth contribution from other ICT technology leaving net boost around 0.35%. The wider literature has focused on the *productivity* potentials of generative AI automation, which are depicted in the graph below, however these do not factor in the additional innovation potentials.

Productivity boost from generative AI
Average annual productivity growth in %



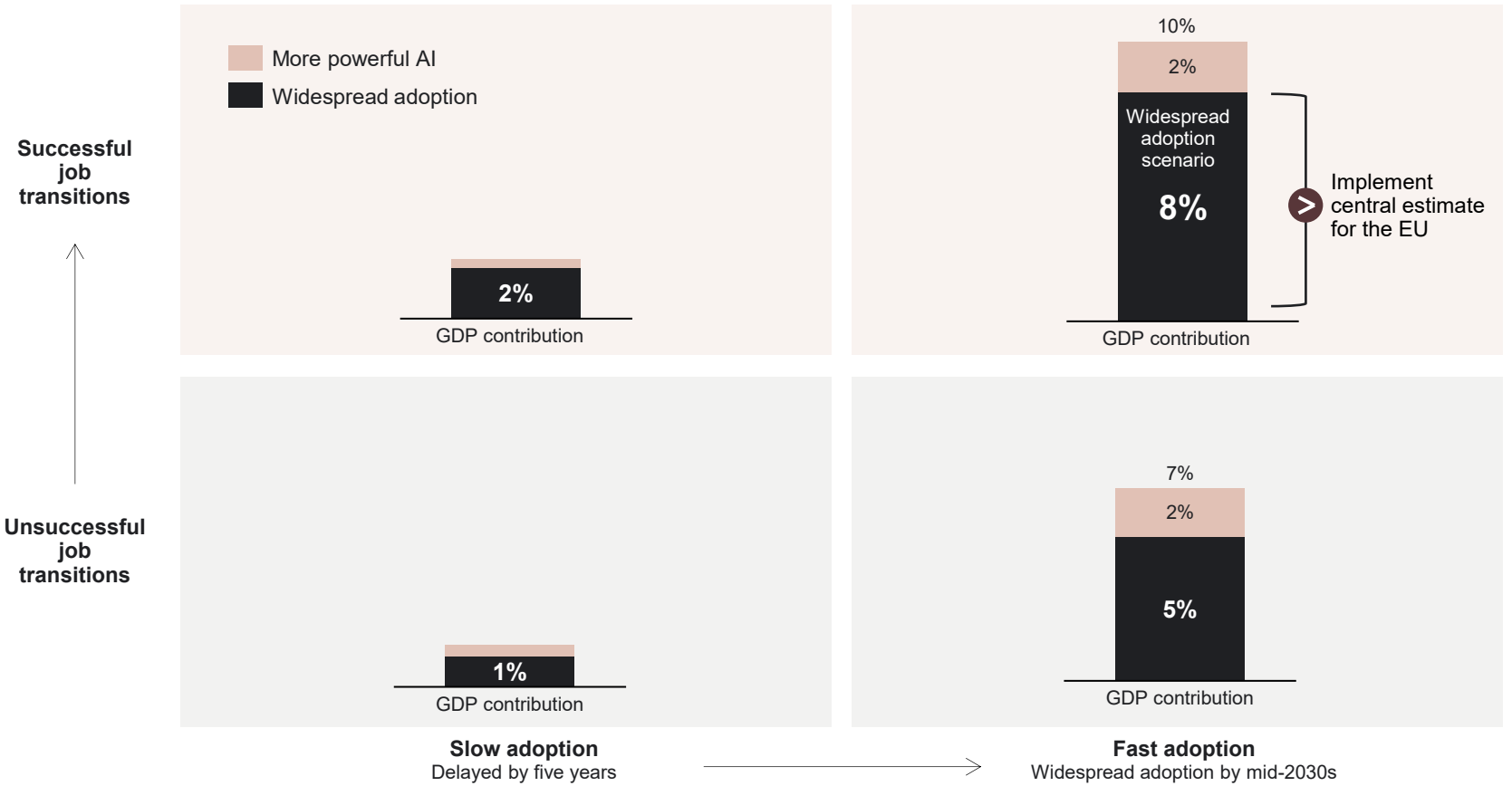
- While our estimate in ICG (2024) is in line with several main studies (Bergeaud 2024, IMF 2025, OECD 2024 and Aghion and Bunel 2024), there are still major differences in the assumptions on how much of the technical (gross) potential Europe will be able to capture.
- Apart from Acemoglu (2024), which includes more modest assumptions, most of the studies reach a similar range of gross boosts from GenAI around 0.7-1.2 p.p., and the main differences come down to how large a potential is economically profitable to realise.
- The IMF (2025) study looks at a five-year time horizon, whereas other studies look at a ten-year period. This also explains why they assume a lower net share of around 18% of the technical potential.
- Bergeaud (2024) provides a range of scenarios with a central estimate of a 0.29 p.p. net annual boost. This highlights the overall uncertainty while still pointing to a steady, positive gain over a decade.
- Aghion (2024) calculates a 0.68 p.p. median gross annual growth over ten years using a task-based framework. The study also notes that historical parallels to past technological revolutions could push the estimate up to 1.3 p.p.
- More details can be found in the annex, p. 37.

Note: Quantifications of the potential of generative AI vary across methodologies, assumptions and economies. *Econometric* studies can quantify existing generative AI productivity boosts but are still dependent on econometric design and a limited empirical database. *Potential* studies, such as Implement's and those shown in the table, quantify the economic benefits from AI from an acknowledged theoretical framework.
Source: Implement Economics based on Acemoglu (2024), IMF (2025), Bergeaud (2024), Briggs and Kodnani (2023), Aghion & Bunel (2024), and Implement Consulting Group (2024).

The size of the adoption potential depends on model capabilities, adoption speed, and job transitions

The potential impact of generative AI on GDP in the EU

% annual contribution from baseline GDP after ten years



- The potential of generative AI is equivalent to 8% of GDP in our main scenario (*widespread adoption*). This assumes productivity gains, re-employment and technology adoption consistent with the assumed adoption curve.
- A five-year delay in capturing the benefits of generative AI is estimated to reduce the GDP growth potential in ten years from 8% to only 2%. The timing and speed of adoption is among the most uncertain elements of the economic potential of AI in the EU.
- If the EU achieves widespread adoption but fails to reskill and effectively reemploy its workforce, the potential impact of generative AI would drop from 8% to 5% of GDP.
- In an ideal scenario where the EU can significantly boost innovation, leading to the development of more powerful AI, while also ensuring widespread adoption and successful job transitions, the economic potential of generative AI could rise to 10% of GDP.
- In addition to these channels, potentially restrictive regulation from the EU could reduce the potential by contributing to uptake delays and to the availability of AI models and tools on the European market.

Note: The “more powerful AI” scenario assumes a productivity boost from generative AI for complemented workers 1.5x larger than baseline assumptions. The “unsuccessful reskilling/re-employment” scenario assumes little or no re-employment of displaced workers and freed-up resources for complemented workers.
Source: Implement economics based on Briggs and Kodnani (2023).

Why TDM access matters for AI capabilities and innovation

TDM for commercial AI training is currently permitted in the EU on an opt-out basis, but this regime is debated in the context of GenAI.

Europe’s copyright regime was designed to strike a balance between encouraging creation of new content and ensuring that others can innovate based on existing content

Balancing for common good

The exceptions for commercial text and data mining are intended to promote overall public welfare.

The balance is important not only for users of knowledge goods, but for creators as well.

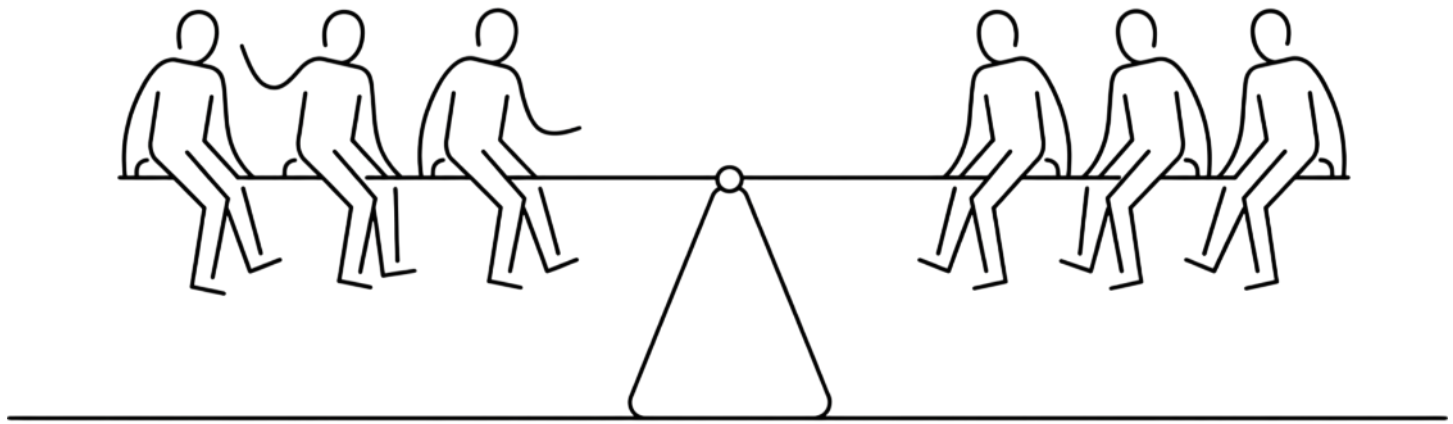
Without the appropriate balance between protection and access, copyright systems risk not only impoverishing the public but may also undermine creative creation over the longer-term.

The TDM regime aims to balance the overall public interest of maximising innovation while maintaining the incentive for continued creation of new content.

The copyright balance

Rewarding creators so they keep creating

Enabling innovation so others can build on what exists



Balancing the considerations of rightsholders…

… With the economic and social benefits

TDM exceptions are linked to stronger AI innovation and commercialisation

Unlocking AI's potential through the availability of training data

AI models require large amounts of training data, much of which may be protected by copyright. Whether developers can legally use this data without permission varies widely across countries, ranging from broad fair use and text and data mining (TDM) exceptions to strict requirements for rightsholder consent.

Research from the Lisbon Council provides a detailed legal analysis of the European copyright regime in relation to AI.

Bruegel research provides the economic arguments in favour of reducing copyright protection for generative AI inputs and outputs.

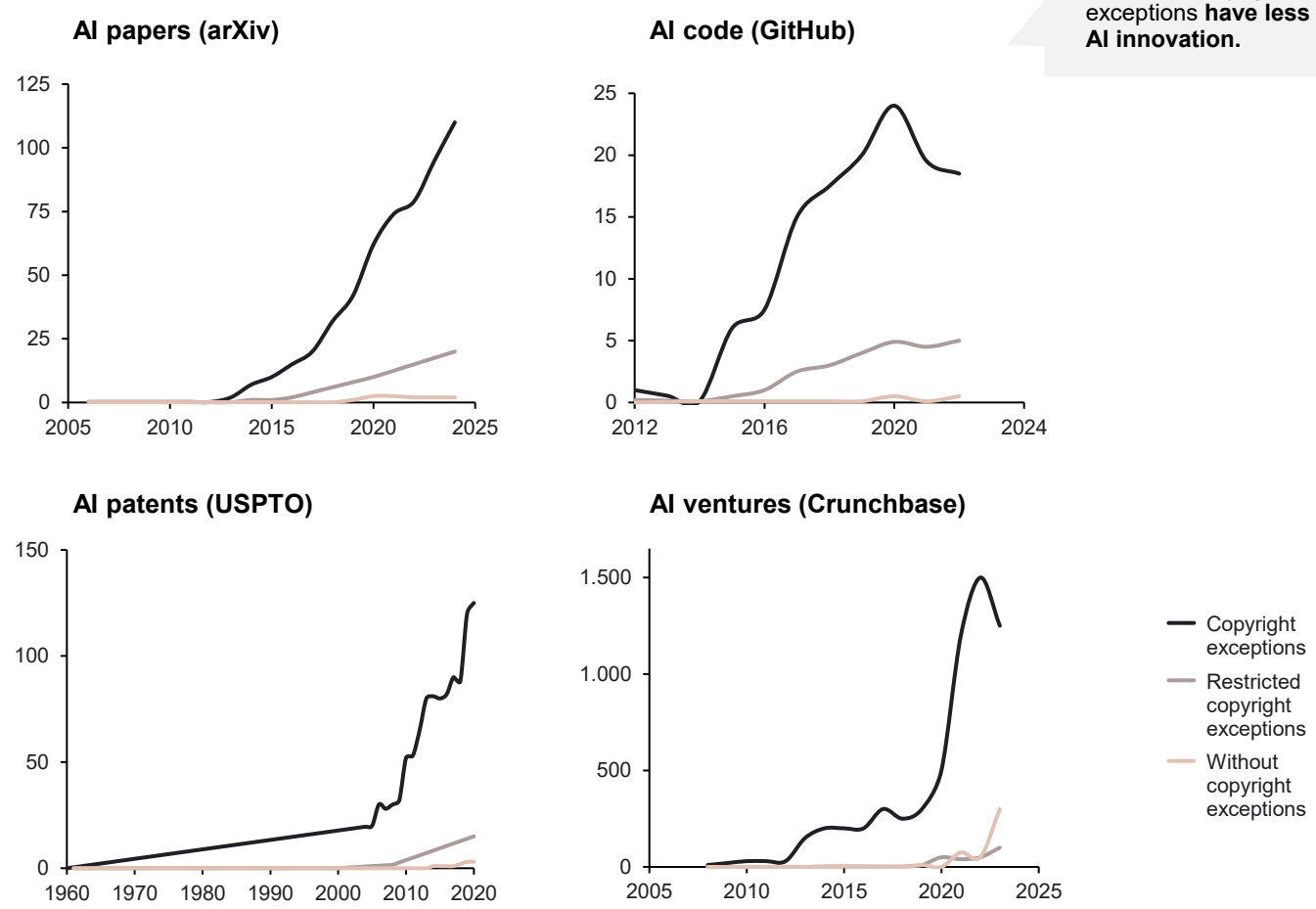
Two recent studies look at how differences in copyright regimes relate to AI innovation, one using cross-country data and the other using economic modelling.

Peukert (2025) examines the link between access to training data, enabled by exceptions in national copyright law, and several measures of AI innovation. Countries are grouped by the breadth of their exceptions, from broad to restricted to none.

Countries with broader exceptions show higher levels of innovation, with faster growth in AI publications, AI code, patents, and ventures. The gap is visible not just in levels but also in the pace of growth during the late 2010s and 2020s.

Yang and Zhang (2024) use a theoretical model to show that abundant training data and unconstrained use of human-generated data protected by copyrights improve GenAI quality.

AI innovation, group averages (Peukert 2025)



Evidence shows that economies with no or restricted copyright exceptions **have less AI innovation.**

Note: Figures adapted and approximated from Peukert (2025).
Source: Implement Economics based on Lisbon Council (2025), Martens (2024), Martens (2025), Peukert (2025) and Yang and Zhang (2024).

Companies across the EU's key industries already rely on the commercial TDM exception to train models and build applications

PHARMACEUTICALS

Literature mining in pharma

What: Accelerating drug discovery and medical research.

How: Pharmaceutical companies, including **Novo Nordisk** and **Sanofi**, use natural language processing to mine millions of biomedical publications, extracting complex associations between genes, diseases, and drug targets. This approach accelerates drug discovery, famously enabling BenevolentAI to identify baricitinib as a COVID-19 treatment.

TRANSLATION SERVICES

Unlocking efficient translation

What: Delivering real-time accurate translation for professionals.

How: DeepL trains neural machine translation to allow individual users and professionals to quickly translate documents. The models are trained on roughly one billion bilingual sentence pairs scraped from corporate websites, news outlets, and other online sources. Large-scale crawling yields diverse languages, domains, and styles.

AUTOMOTIVE

Powering in-vehicle voice assistants with fine-tuned LLMs

What: Powering in-vehicle virtual assistants and diagnostic systems.

How: Volkswagen integrates a fine-tuned LLM-based assistant via **Cerence**, making ChatGPT standard in 2024 and answering about 10,000 vehicle- and brand-specific operational questions. Cerence's product answers questions based on OEM-verified data and owner manuals.

FINANCIAL SERVICES

Detecting ESG and reputational risk in real-time

What: Identifying ESG and market risk signals in real-time.

How: SEAMm mines large-scale web text to surface controversies and sentiment signals on public and private companies. Its TextReveal platform analyses over 20 billion articles and messages in more than 100 languages from 4 million+ sources, classifying mentions across roughly 90 risk categories.

FASHION

Forecasting trends in the fashion industry

What: Analysing publicly accessible social posts to detect granular fashion trends.

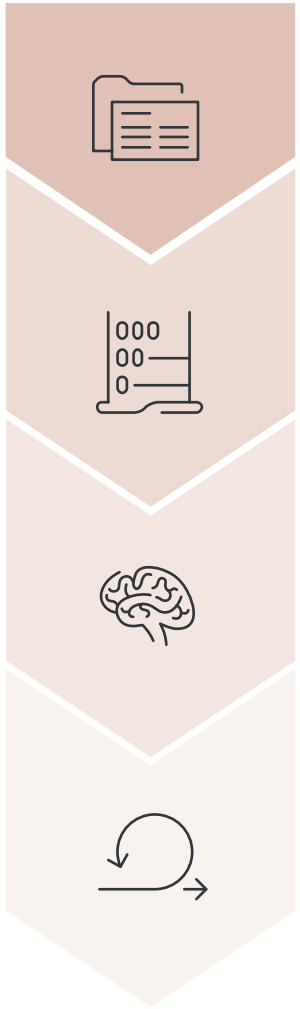
How: Heuritech analyses publicly accessible social media accounts at scale, monitoring around 50,000 public Instagram accounts per geography and 400,000 Weibo accounts. Its AI scans millions of images for 3,000+ fashion details and turns them into aggregated trend forecasts, helping brands such as LVMH, Prada and Adidas anticipate trends.

How these use cases rely on the commercial TDM exception: The EU TDM exception enables innovators to draw on large, multilingual and geographically dispersed corpora, including news articles, scientific publications, social media posts, user-generated images, technical documentation and open web text. It avoids the need to clear fragmented rights, obtain collective licenses or negotiate individual licenses across millions of rightsholders.

Note: In sectors - such as biomedical publications - where a licensing market exists, limiting innovators to a licensed subset of available data would lead to coverage gaps, risking that crucial relationships between data points go undetected. Use cases are based on research by Brinkhof.
Source: Implement Economics based on legal analysis by Brinkhof.

Text and data mining (TDM) enables AI to be trained on real-world data

The technical side of TDM



Data ingestion
The process begins by gathering data to train the model. For text-based models, this can amount to billions of data points from public web domains, open-source repositories, and digital libraries, sometimes enriched with specialised content acquired through voluntary licensing deals and partnerships.

Pre-processing and vectorisation
Raw data must first be converted into a numerical format. For text, words or tokens become high-dimensional numerical vectors through a process called *embedding*. For images, they become arrays of pixel values. The model does not "see" words or "view" images; it only processes numerical representations.

Model architecture
The model is an assembly of complex mathematical functions, often a neural network composed of many layers of interconnected nodes or neurons. These millions or billions of parameters will be tuned during training, starting from random values, and the model does not retain any training data.

The training loop
The model processes data in batches: it takes an input (e.g. a sentence) and makes a prediction (e.g. the next word). A loss function calculates how accurate the prediction was. The model then adjusts its internal weights to reduce this error, with the goal of minimising the loss across the entire dataset. The same logic applies in fine-tuning, where models are adapted using targeted datasets, while retrieval-augmented-generation (RAG) allows models to complement training by retrieving relevant external information at the point of use.

The legal side of TDM

Directive (EU) 2019/790 on Copyright in the Digital Single Market ("CDSM Directive") introduced exceptions for text and data mining ("TDM"). TDM is defined as “any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations”.

Under Article 4, Member States shall provide for an exception regarding copyright and related rights and database rights allowing anyone to reproduce protected works for the purposes of text and data mining, provided that the works are lawfully accessible. The exception is subject to reservations of rights, or opt-outs, by rightsholders, who must express such reservations “in an appropriate manner”.

Recital 18 of the CDSM Directive clarifies the rationale behind Article 4. It emphasises that TDM is not limited to scientific research but is widely used by public and private actors for purposes such as public services, business decision-making and the development of new technologies. At the same time, it recognises legal uncertainty under the existing regime. Article 4 is therefore intended to provide legal certainty and to support innovation, including in the private sector.

The current policy debate

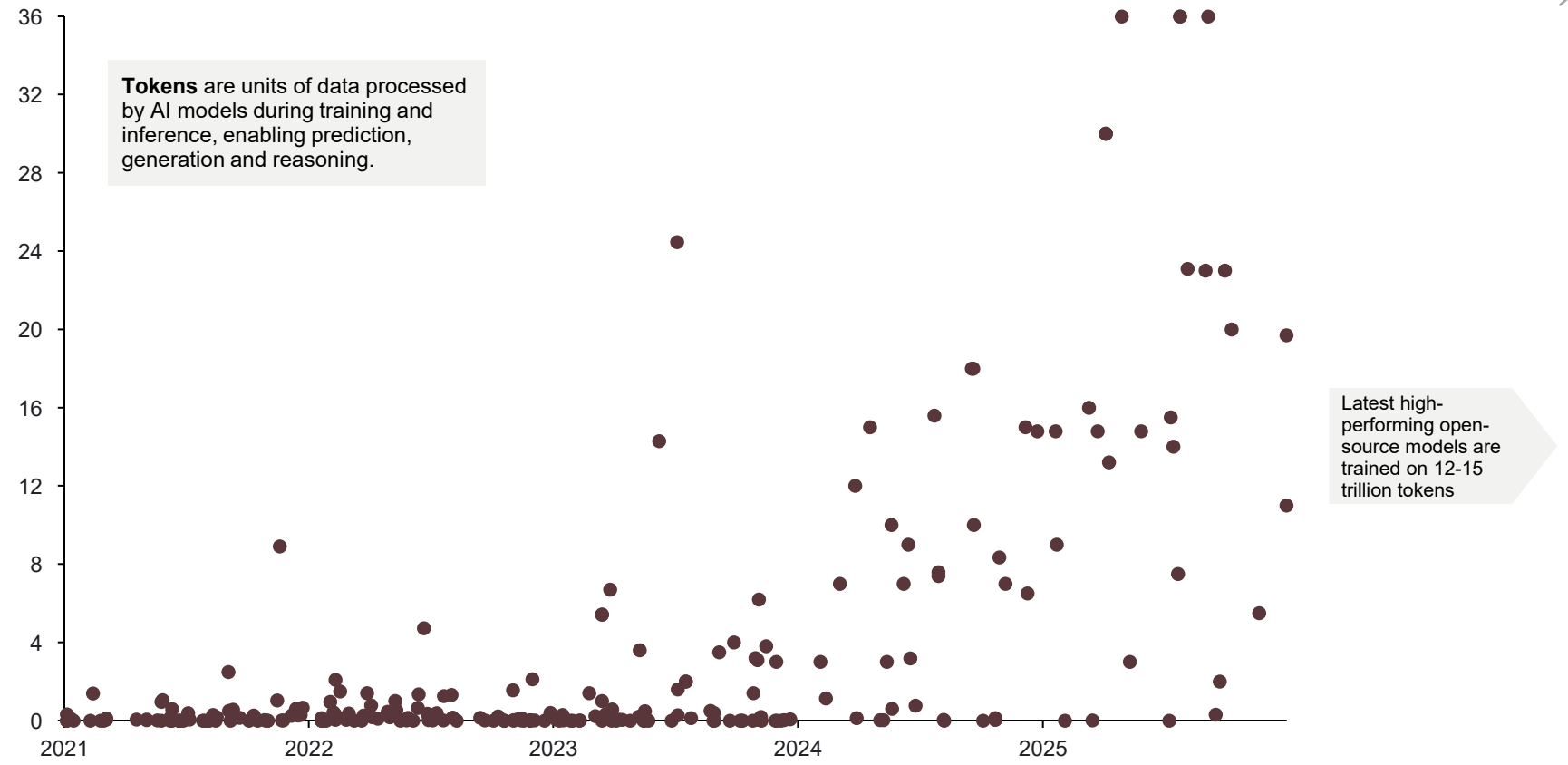
The policy debate around TDM for AI in the EU is evolving. The Artificial Intelligence Act, adopted in 2024, confirms that the TDM exception extends to GenAI training.[1] Recital 105 AIA notes that such models require access to vast amounts of text, images, videos and other data, and that TDM techniques may be used extensively for the retrieval and analysis of such content. Article 53(1)(c) of the AI Act requires providers of general-purpose AI models to draw up a policy to comply with EU copyright law, including methods to identify and comply with opt-outs under Article 4 of the CDSM Directive. The copyright chapter of the Code of Practice for general-purpose AI models further presupposes that GenAI training is covered by the TDM exception.

Many rightsholders and some scholars and politicians argue that the commercial exception should not cover GenAI training.[2] They argue that GenAI often circumvents traditional licensing mechanisms, risks systematic interference with exploitation markets, especially in music, visual art, and journalism, and express concerns over the opt-out mechanism. Policy suggestions include more transparency requirements, mandatory licensing schemes, changes in opt-out mechanisms and requirements on AI models trained outside the EU (extraterritoriality), as analysed further in Lisbon Council (2025).

Note: 1) In 2023 Thierry Breton confirmed that the CDSM Directive’s TDM exceptions “*provide balance between the protection of rightsholders including artists and the facilitation of TDM, including by AI developers*”, further supporting that Article 4 was intended to enable TDM by AI developers. 2) See for example views expressed at the Workshop on Generative AI & Copyright organised by the European Parliament in July 2025.
Source: Implement Economics based on European Parliament, Answer to Parliamentary Question E-000479/2023 and legal analysis by Brinkhof.

While the amount of data used in training of frontier AI models has increased dramatically, most recent developments suggest that diverse data also matters

Number of unique data points used to train AI models
Trillion tokens



- The volume of data used to train frontier AI models grew rapidly after 2023 and the largest models were trained on more than 35 trillion tokens.
- Prior to 2023, the most ambitious models rarely exceeded 5 trillion tokens.
- The increase during the 2023-25 period was driven by a pursuit to explore how model performance would improve by the size of the training data.
- Since then, improved data curation techniques and new training techniques suggest that frontier model performance can also be achieved with around 12-15 trillion tokens, and training on 5 trillion "perfect" tokens is now proven to outperform training on 30 trillion "noisy" ones.
- So, while the surge in training data size may not continue upwards, AI training and fine-tuning still requires massive and diverse datasets.

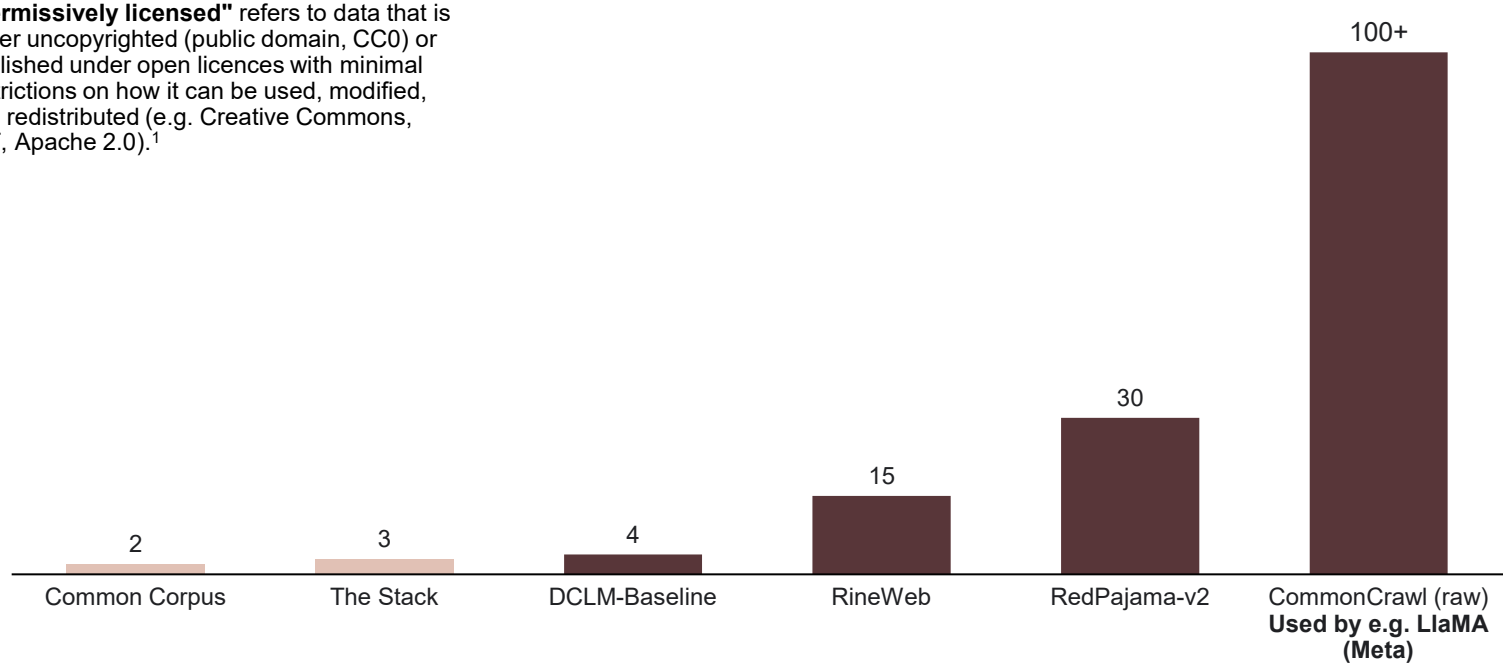
Note: Outliers removed for visual purposes.
Source: Implement Economics based on Epoch AI (2025).

Permissively licensed datasets represent a fraction of the data typically used to train frontier AI models

Size of AI training datasets
Trillion tokens

Other training data sets
Permissively licensed (copyright-conscious)

"Permissively licensed" refers to data that is either uncopyrighted (public domain, CC0) or published under open licences with minimal restrictions on how it can be used, modified, and redistributed (e.g. Creative Commons, MIT, Apache 2.0).[1]

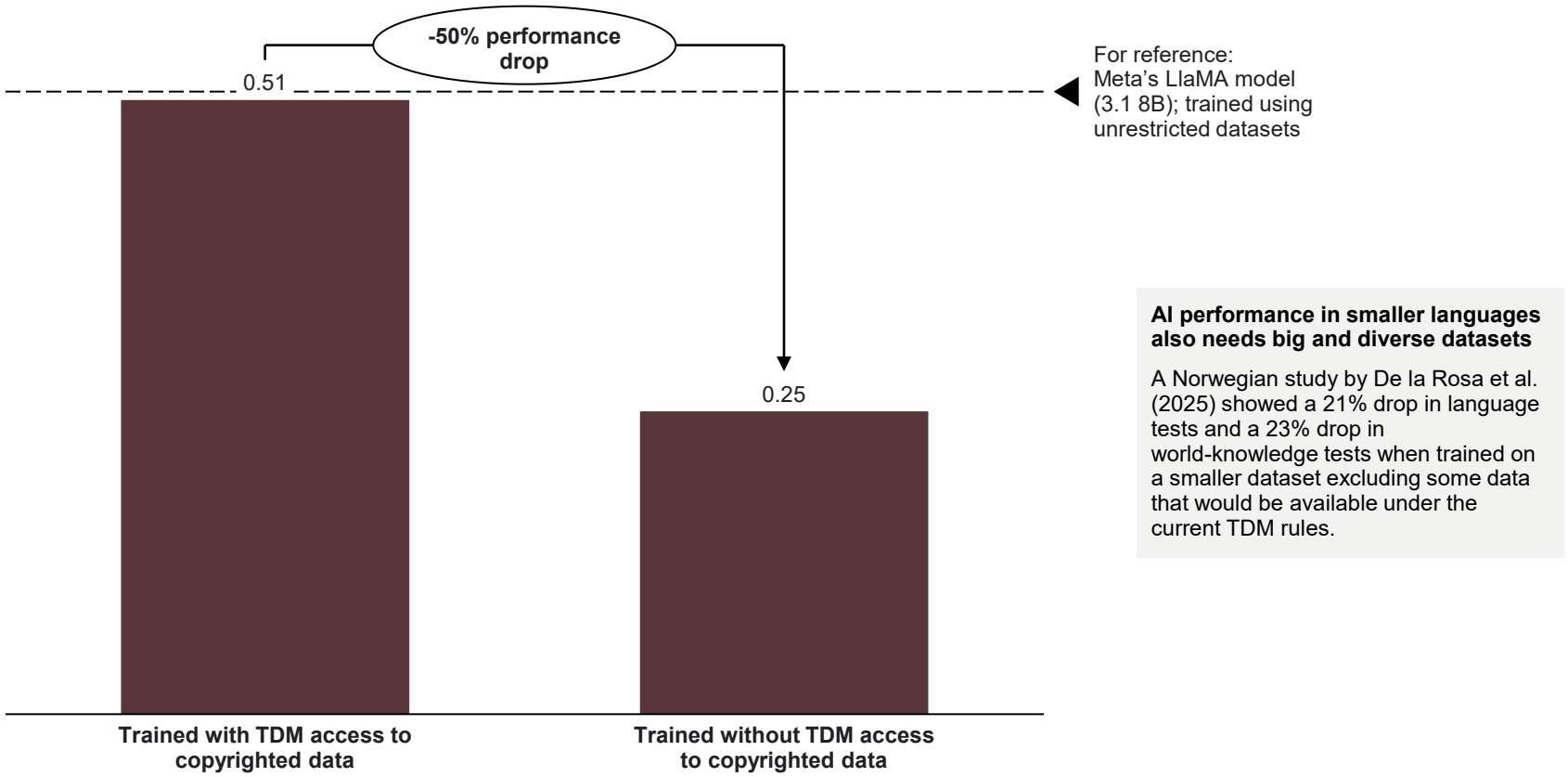


- The largest permissively licensed training dataset currently available contains approximately 3 trillion tokens, roughly 33 times smaller than CommonCrawl's raw dataset of at least 100 trillion tokens. Most frontier models had already surpassed the 3 trillion token datasets by 2023.
- This gap exists because permissively licensed datasets must exclude large categories of commercial content, including news articles, books, and academic publications, that are central to developing strong language understanding and factual reasoning capabilities.
- The shortfall is particularly acute for low-resource European languages, where the pool of freely available, non-copyrighted text is too small to train competitive models without supplementary data from copyrighted sources (see page 17).
- Access to European training data is key to one of the objectives of the AI Act of developing AI models in line with European values.[2.]
- While rightsholders can already opt-out under the current commercial TDM regime, a shift to opt-in or a mandatory licensing regime could reduce the training data available to European model and application developers.
- For state-of-the-art AI training and fine-tuning to flourish in Europe, innovators need a European common data space with size and diversity.

Note: 1) Permissively licensed datasets include only content that is uncopyrighted, in the public domain, or published under open licences (e.g. Creative Commons, MIT, Apache 2.0). Token counts reflect raw or total reported size at time of release and may differ from deduplicated or filtered training-ready versions. 2) See e.g. the AI Act recital 8 mentioning that "…rules should be clear and robust in protecting fundamental rights, supportive of new innovative solutions, enabling a European ecosystem of public and private actors creating AI systems in line with Union values."
Source: Implement Economics based on publicly available dataset documentation from Hugging Face, CommonCrawl Foundation, and respective dataset maintainers (BigCode Project, Pleias, Together AI, Allen AI).

GenAI models are half as good at complex reasoning when trained on smaller datasets

Model performance on complex reasoning
Benchmark test score (max = 1)



AI performance in smaller languages also needs big and diverse datasets

A Norwegian study by De la Rosa et al. (2025) showed a 21% drop in language tests and a 23% drop in world-knowledge tests when trained on a smaller dataset excluding some data that would be available under the current TDM rules.

- When AI models are trained on too small datasets, it results in weaker model capabilities.
- A large-scale Swiss scientific research project, Apertus (2025), performed supervised fine-tuning (also called post-training) on the same model using different training datasets.
- They found that the same model, when post-trained without access to copyrighted data, suffered a 50% drop in performance on complex reasoning.
- More precisely, the Swiss researchers note that adding license filtering results in "*particularly severe drops on MMLU chain-of-thought evaluation from 0.513 to 0.253 (a 51% decrease)*", cf. Apertus (2025).
- While the Swiss study shows severe degradation in performance for complex reasoning, it also demonstrated that filtering away licensed data from post-training produced virtually no performance degradation for more rudimentary tasks.
- Another study by De la Rosa et al. (2025) showed that access to large and diverse training and post-training data is also important for AI performance in smaller European languages.

Note: Performance measured on MMLU CoT-strict, a general knowledge and complex reasoning benchmark. Both bars use the same base model (Apertus 8B at 10T tokens), fine-tuned on different data configurations using the Tulu3 post-training mixture. The left bar shows Tulu3 with decontamination only (score: 0.51), while the right bar shows Tulu3 with both decontamination and license filtering applied (score: 0.25). A separate openly licensed dataset (OLMo2) produced a similarly low score of 0.33, suggesting the performance drop is not unique to one dataset.
Source: Implement Economics based on Apertus V1 Technical Report (2025) and De la Rosa et al. (2025)

The cost of getting it wrong

Restricting commercial TDM in the EU could put a significant share of the economic potential of AI at risk.

The economic potential from AI is of considerable importance to Europe's economy, and the value at risk from changes in the commercial TDM regime should be considered

The current commercial TDM regime is challenged by suggestions to introduce…

A mandatory licensing regime for AI training, rather than voluntary market-driven licensing, would alter the current commercial TDM framework and add transaction costs and legal friction at scale.

Disproportionate transparency requirements, whereby AI developers would have to publish granular details of the data used to train their models, potentially exposing trade secrets and creating security vulnerabilities.

Designation of new opt-out mechanisms that are fragmented, unworkable, and lack widespread adoption.

Potential extraterritorial effects of EU requirements, whereby models trained outside the EU may still need to comply with EU licensing, transparency, or opt-out rules when placed on the EU market, even though they already took steps to comply with copyright (including through deals) in their domestic market.

Affecting the AI economic opportunity in Europe…

Adoption speed

Slower adoption and diffusion
The commercial TDM exception is central to developing European AI applications and models. When new regulation is debated it creates uncertainty about future market conditions. This can cause European AI developers to pause or delay the development of new AI applications. With a delay in developing applications comes a risk of delay of adopting AI, since AI adoption requires AI applications.

Model capabilities

Less powerful AI
Commercial text and data mining is required for training and post-training fine-tuning. Restricting the commercial TDM access to large and diverse data can reduce model performance which in turn will cause harm to:
a) European AI developers' ability to train new frontier models and
b) European companies' ability to create domain-specific AI systems using post-training fine-tuning

AI innovation

Lost R&D potential
AI-driven R&D efficiency gains depend directly on frontier model capabilities, which rely on broad TDM access. Restricting this access limits the performance of models used across research workflows, from drug discovery to materials science. European researchers would lose a competitive edge if their AI tools lag behind those in jurisdictions with fewer data access constraints.

AI value chain

Lost development potential
TDM is critical to training and fine-tuning European models and applications. EU-specific requirements can also reduce the availability of frontier models developed abroad if providers choose not to place their newest systems on the EU market. Domestic development of foundation models and services depends on access to massive amounts of training data. A supportive commercial TDM regime helps anchor AI production, talent, and venture capital within the EU rather than shifting activity to more permissive jurisdictions.

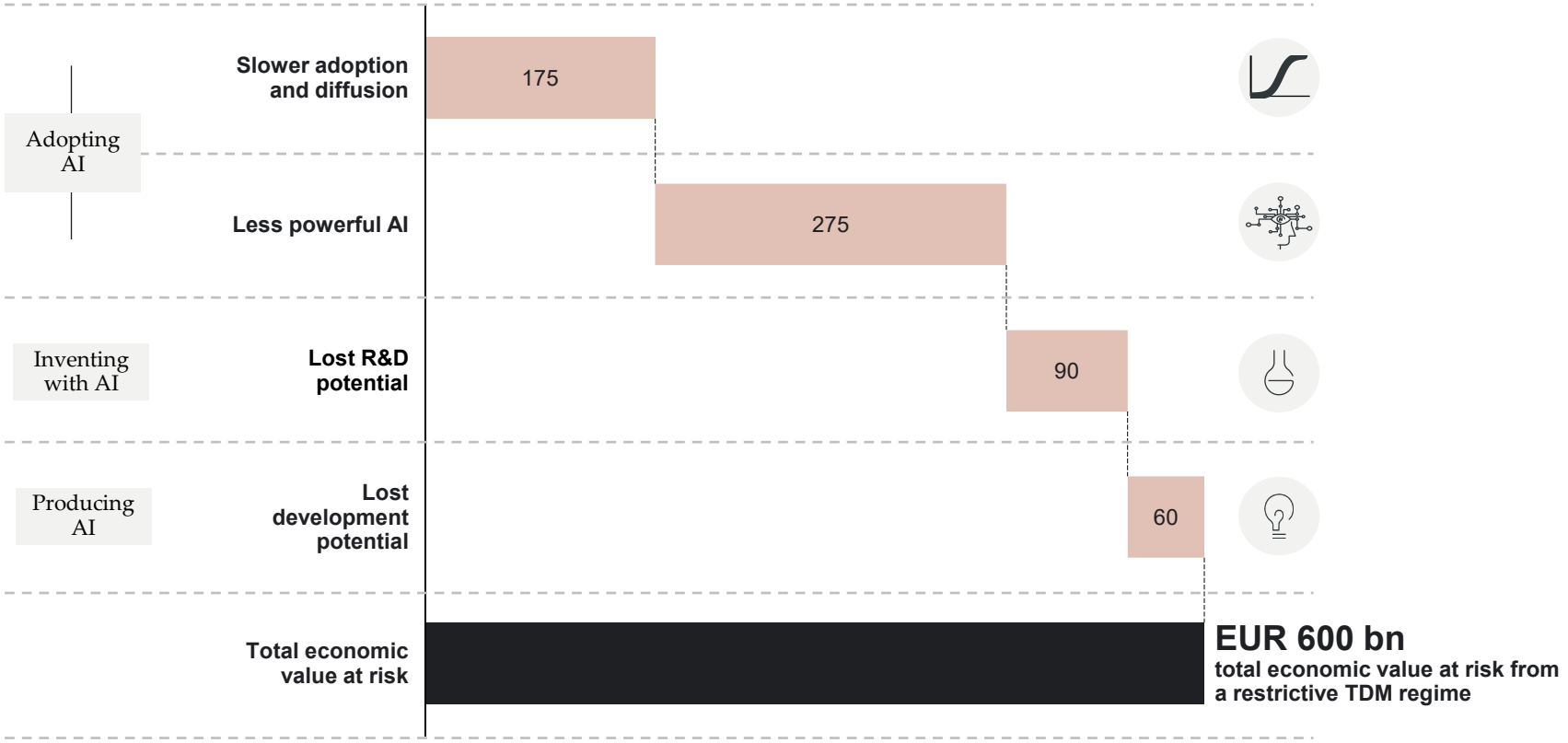
These effects are mutually interdependent

Note: Current EU discussions also concern machine-readable opt-out protocols within the existing opt-out regime; the risk arises if endorsed mechanisms are fragmented or not widely adopted.

Restricting commercial TDM could put EUR 600 billion in economic value at risk across four mutually reinforcing channels

Economic value at risk

EUR billion



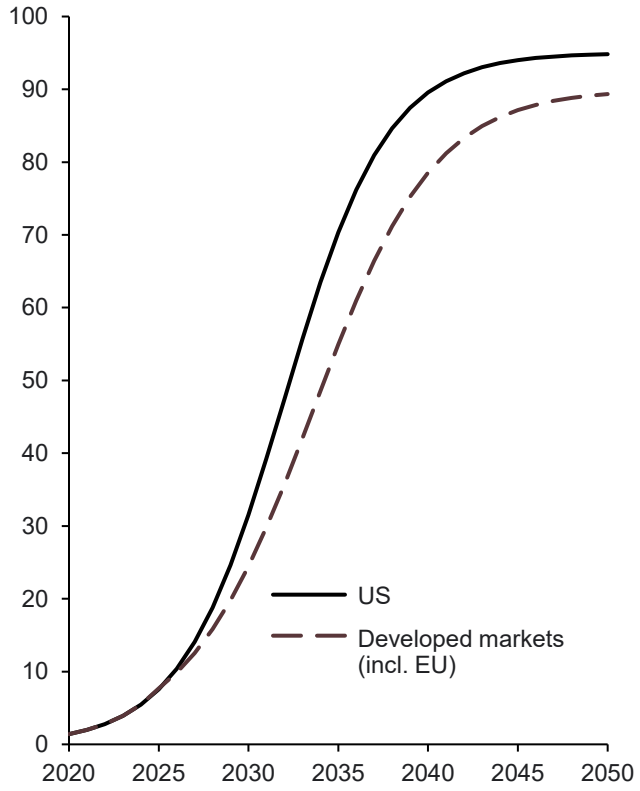
- A restrictive commercial TDM regime is estimated to put EUR 600 billion of annual economic value at risk across four channels.
- **Slower adoption and diffusion**: Compliance friction and licensing complexity delay enterprise deployment. The model assumes a conservative one-year adoption lag, which alone accounts for EUR 175 billion in lost potential.
- **Less powerful models**: Filtering copyrighted content from training data, to the extent this is possible, weakens AI models. If the EU does not have access to the most powerful AI models, EUR 275 billion is put at risk.
- **Lost R&D potential**: Research-intensive sectors depend on frontier model capabilities for R&D efficiency gains. Weaker models and adoption delays together put around EUR 90 billion at risk.
- **Lost development potential**: European model developers and application builders need large-scale TDM access to compete globally. Without it, the EU's producing potential drops from EUR 200 billion to roughly EUR 140 billion.
- These channels are reinforcing. Weaker models slow adoption because lower accuracy, reliability, and task automatability reduce the case for deploying AI. Slower adoption affects the case for domestic development, and a smaller ecosystem limits innovation gains further.

Source: Implement Economics based on O*NET, Apertus (2025), Briggs and Kodnani (2023).

Regulation and the uncertainty around regulation cause delays in the diffusion and adoption of technologies in the EU

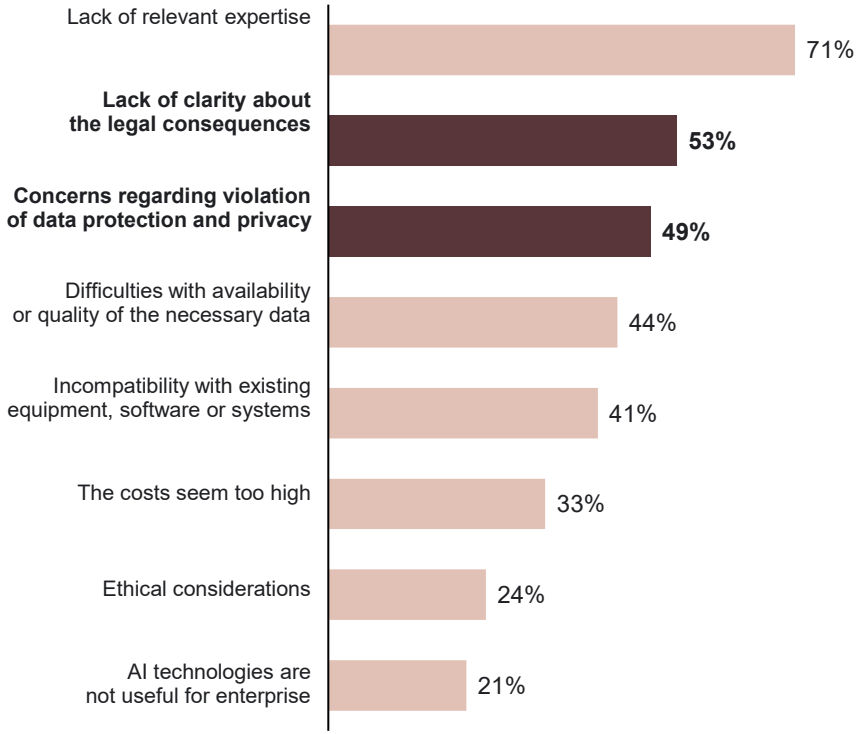
The EU is on a slower and shallower AI adoption trajectory due to regulatory frictions …

Adoption of AI
%



… And regulatory uncertainty ranks among the top barriers to AI adoption for European enterprises

Reason for not using AI technologies among EU enterprises
% of enterprises



Slower adoption and diffusion

- The EU's AI adoption trajectory is already slower and shallower than that of the US because existing regulatory frictions are delaying diffusion, but these delays have so far taken place within a stable copyright framework.
- Regulatory friction is already producing measurable delays in the availability of major AI products in Europe, with lags ranging from 4 months to more than 12 months. That shortens the period in which EU firms and consumers can benefit from frontier tools, and some delays appear to reflect both direct enforcement and firms holding products back because of legal uncertainty.
- Regulatory uncertainty is a key driver of Europe's slower AI uptake. More than half of EU enterprises, 53%, cite unclear legal consequences as a reason for not adopting AI, while 49% cite data protection and privacy concerns, making regulation one of the clearest barriers to adoption beyond the skills gap.
- Changes to the commercial TDM regime would likely introduce a new and larger source of friction on top of existing ones, further delaying access to frontier AI models and tools. A comparable UK analysis argues that a restrictive TDM regime would lead to a persistent 5% lower adoption level relative to the EU baseline.

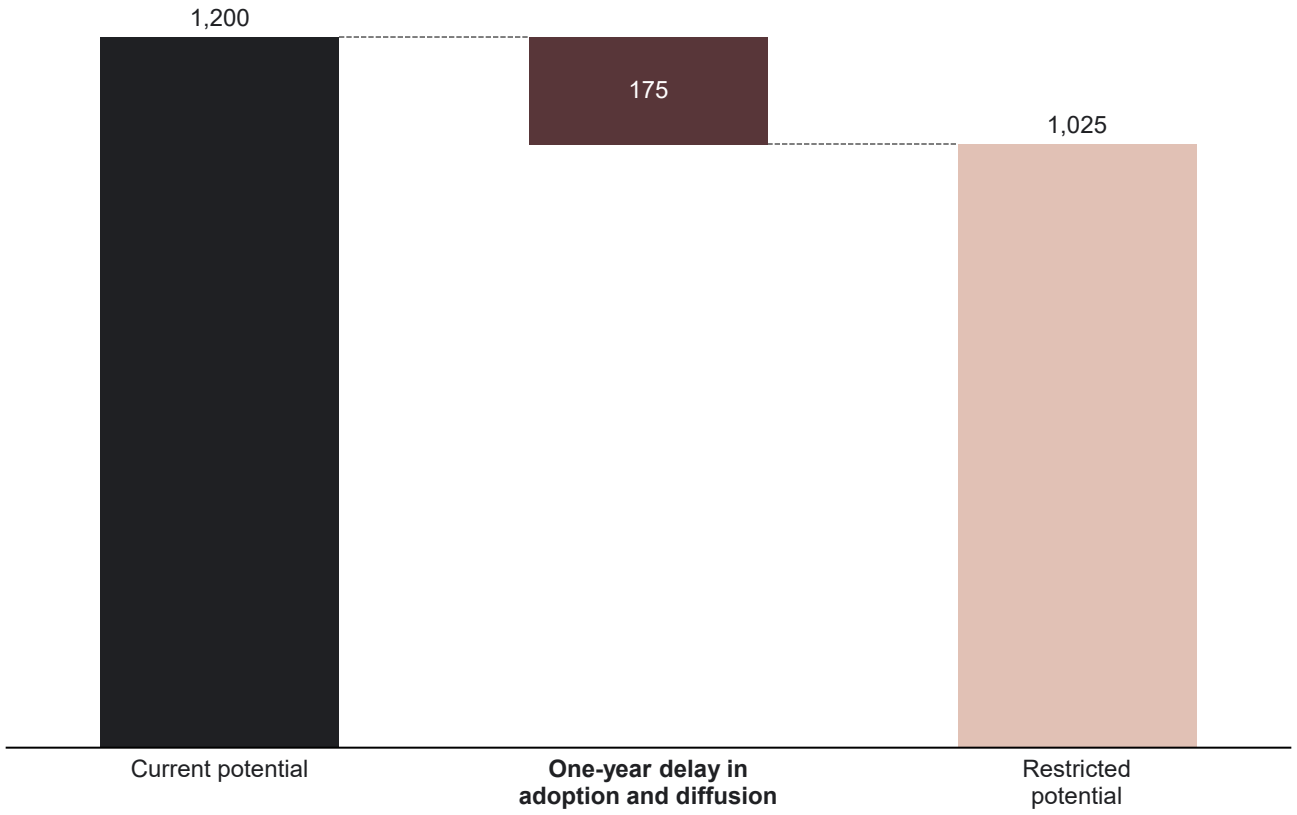
Source: Implement Economics based on Briggs and Kodnani (2023), Centre for British Progress (2025), and Eurostat.

Restrictions on commercial TDM can cause delays to the diffusion and firm adoption of AI – a one-year delay would put EUR 175 billion of the adoption potential at risk

Slower adoption and diffusion

The annual AI adoption potential at widespread adoption

EUR billion increase from baseline GDP after a ten-year adoption period



- A restrictive commercial TDM regime is estimated to put EUR 175 billion of the EU's annual AI adoption potential at risk through slower uptake alone.
- The commercial TDM exception is central to developing European AI applications and models. When new regulation is debated, it creates uncertainty about the future regulatory landscape, causing European AI developers to pause or delay the development of new applications. Since AI adoption depends on applications being available in the first place, delays in development translate directly into delayed adoption.
- We have modelled a one-year lag in AI adoption relative to the status quo workable regime.
- In a similar study, the Centre for British Progress (2025) finds that even a 5% annual adoption lag from opt-in requirements carries billions in lost economic value over a five-year horizon in the UK.
- Some of the potential effects of delays may already have been incurred as a result of the signalling around potential changes from public EU institutions.

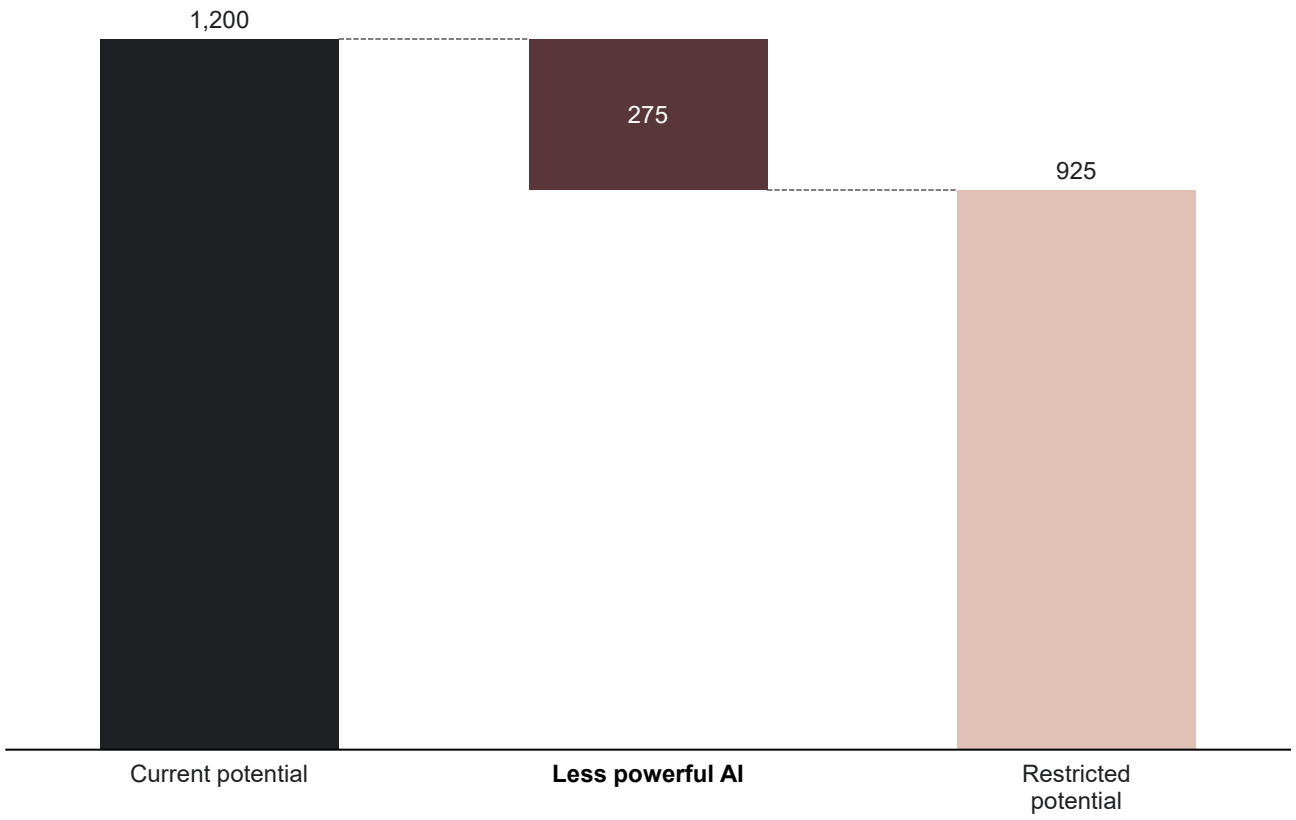
Source: Implement Economics based on O*NET, Centre for British Progress (2025), Apertus (2025), and Eurostat.

Restrictions on commercial TDM could cause a decrease in the model capabilities available in the EU, putting an additional EUR 275 billion of the adoption potential at risk

Less powerful AI

The annual AI adoption potential at widespread adoption

EUR billion increase from baseline GDP after a ten-year adoption period



Actual losses could be higher

Advanced AI models increasingly rely on real-time web searching (retrieval-augmented generation, or RAG) to enhance their accuracy and performance through so-called *grounding*. This live data retrieval acts as another form of text and data mining.

The EUR 275 billion estimate does not include the value of this live web access, meaning the true economic losses from strict restrictions that impact this aspect of frontier models would likely exceed those shown.

- Restrictions on training data access are estimated to reduce the EU's annual AI adoption potential by EUR 275 billion due to less capable models.
- Filtering out licensed or copyrighted content from training data, to the extent this is possible, measurably degrades what models can do. These capability losses matter because they directly reduce the share of work activities that can be meaningfully augmented by AI. Applied across all occupations in the EU, even a modest decline in model performance narrows the productivity boost available to each worker.
- The effect is not uniform across tasks. Performance drops are concentrated in step by step reasoning and broad academic knowledge, precisely the capabilities most valuable for high-productivity sectors like healthcare, legal services, and financial analysis.
- Without access to diverse training data, the EU risks landing at a restricted potential of EUR 925 billion, roughly 20% below what would be achievable under the current TDM regime.
- The current commercial TDM regime with an opt-out mechanism already supports the full adoption potential. However, if opt-out protocols remain poorly standardised, or new, immature protocols are inconsistently enforced, a smaller share of that potential could be at risk due to gaps in its practical implementation.

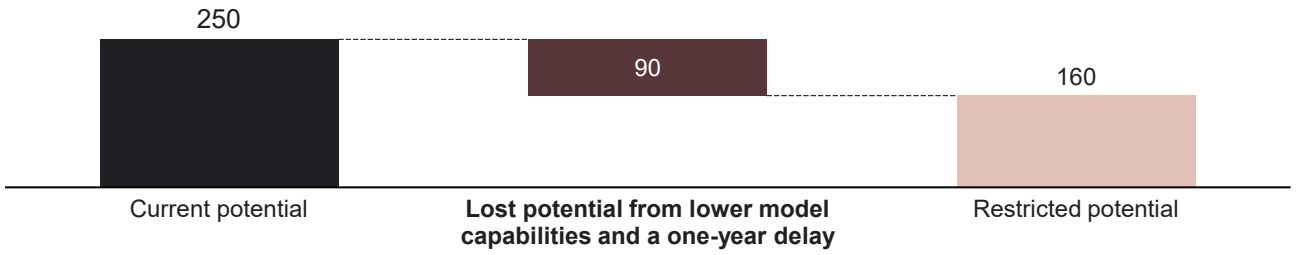
Source: Implement Economics based on O*NET, Apertus (2025), and Eurostat.

R&D applications of AI rely on frontier model capabilities, and commercial TDM restrictions would reduce this potential by EUR 90 billion

Lost R&D potential

Annual R&D AI potential at widespread adoption

EUR billion increase from baseline GDP after a ten-year adoption period



- A restricted TDM regime is estimated to put a third of Europe's annual inventing with AI potential at risk. This structural bottleneck could decrease the projected baseline potential of EUR 250 billion down to a restricted realisation of EUR 160 billion.
- Research-intensive sectors rely heavily on frontier AI models for R&D and on fine-tuning of domain-specific AI applications.
- Limiting licensed data degrades complex reasoning while legal frictions delay enterprise adoption. Together, these effects erode AI-driven productivity and innovation by an estimated EUR 90 billion.
- Historical evidence highlights the severe risk of such legal uncertainty. Strict copyright regimes have historically driven European researchers to abandon TDM projects or relocate them to the US.[1.]
- Failing to maintain a workable commercial TDM regime threatens to disadvantage the EU's research-heavy sectors. Ensuring legal certainty is critical to unlocking the estimated 10-20% efficiency boosts in scientific domains and maintaining Europe's global competitiveness in future technological innovation.[2.]

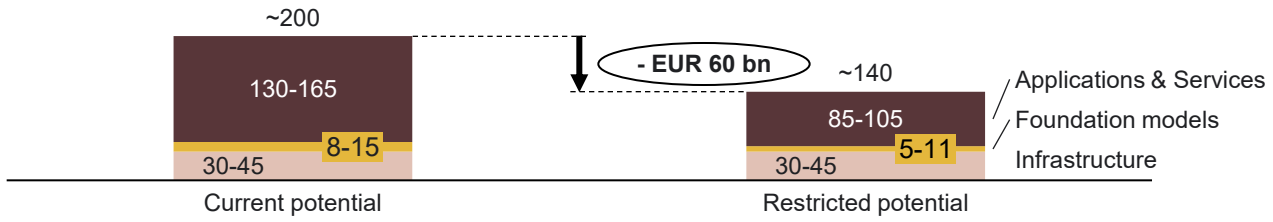
Note: 1) Filippov (2014) analyses the impact of opt-in copyright regimes and finds that legal uncertainty historically caused European researchers to abandon TDM projects and move them to the US; 2) McKinsey (2023) find 10-15% productivity gains from AI delivered as percentage of overall R&D costs. Babina et al (2021) find 18-20% increase in sales due to AI adoption through product innovation.
Source: Implement Economics based on O*NET, Apertus (2025), Briggs and Kodnani (2023), Filippov (2014), McKinsey (2023), Babina et al. (2024), and Eurostat

Restricting commercial TDM would also jeopardise EUR 60 billion in potential value creation from European foundation models and AI applications & services

Lost development potential

Annual producing AI potential at widespread adoption

EUR billion increase from baseline GDP after a ten-year adoption period



TDM restrictions would reduce the possibilities for European developers to produce foundation models and AI applications and services that are competitive with global alternatives developed under less strict conditions.

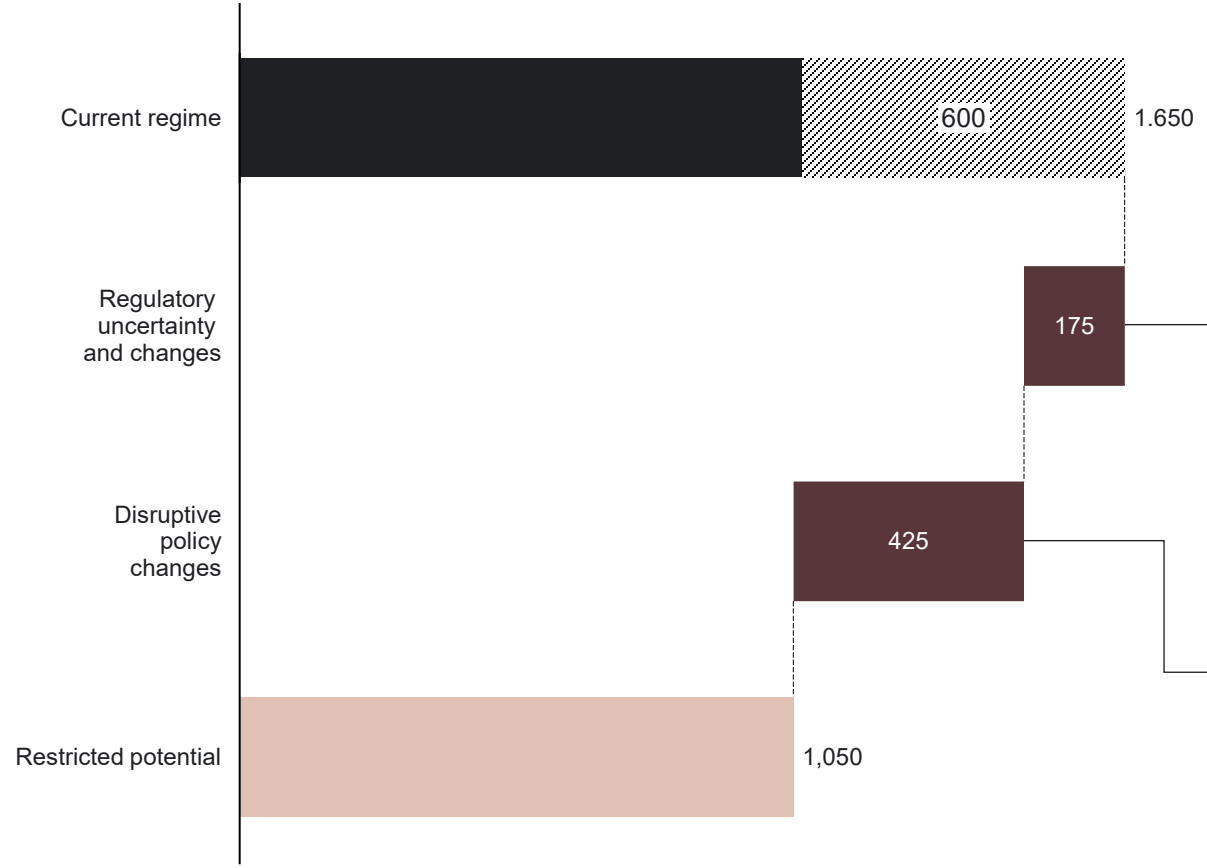
- Europe's domestic AI capacity is highly vulnerable to copyright frictions. A restricted commercial TDM regime would put around EUR 60 bn of annual potential at risk, reducing total potential from roughly EUR 200 bn to EUR 140 bn.
- This reduction reflects the data needed to train competitive foundation models. With less access to training data than developers in more permissive jurisdictions such as the US, Japan or Singapore, European foundation model potential would fall from EUR 8-15 bn to EUR 5-11 bn.
- The same effect extends up the value chain. Applications and services potential would fall from EUR 130-165 bn to EUR 85-105 bn, since these products depend on both strong base models and commercial TDM exceptions for fine-tuning.
- Without permissive rules for essential tasks like model fine-tuning, EU developers are likely to fall behind global alternatives, undermining the EU's own ambition to build competitive domestic AI capabilities.
- Ultimately, expanding the domestic AI value chain requires workable data policies. Failure to provide this legal certainty would compromise Europe's position as an active producer in the global AI market, leaving commercial TDM restrictions in contradiction with the EU's broader industrial objective of strengthening Europe's AI capacity and technological position.

Note: Radeisen (2026) analyses the implications of TDM restrictions on the development of foundation models and states that TDM is fundamental for this training process. Peukert (2025) shows that jurisdictions with flexible copyright regimes successfully commercialise far more AI applications, evidenced by significantly higher rates of patent filings and venture creation. Infrastructure spending, estimated at EUR 30-45 bn, is assumed to remain unchanged regardless of the TDM regime, as data centre and compute investments are less directly tied to copyright restrictions on training data.
Source: Implement Economics based on O*NET, Peukert (2025), Radeisen (2026), and Eurostat.

Reductions in the EU's economic potential from AI are differently affected by the initiatives being discussed in the EU

Economic potential of AI in the EU at widespread adoption

EUR billion



Changes to TDM regime that can put this value at risk…

Regulatory uncertainty around TDM may already have contributed to delaying AI development and adoption across the EU. When new regulation is debated, it can create uncertainty about the future regulatory landscape, causing European AI developers to pause or delay the development of new applications. Since AI adoption depends on applications being available in the first place, these delays may translate directly into slower adoption, putting EUR 175 billion of annual potential at risk.

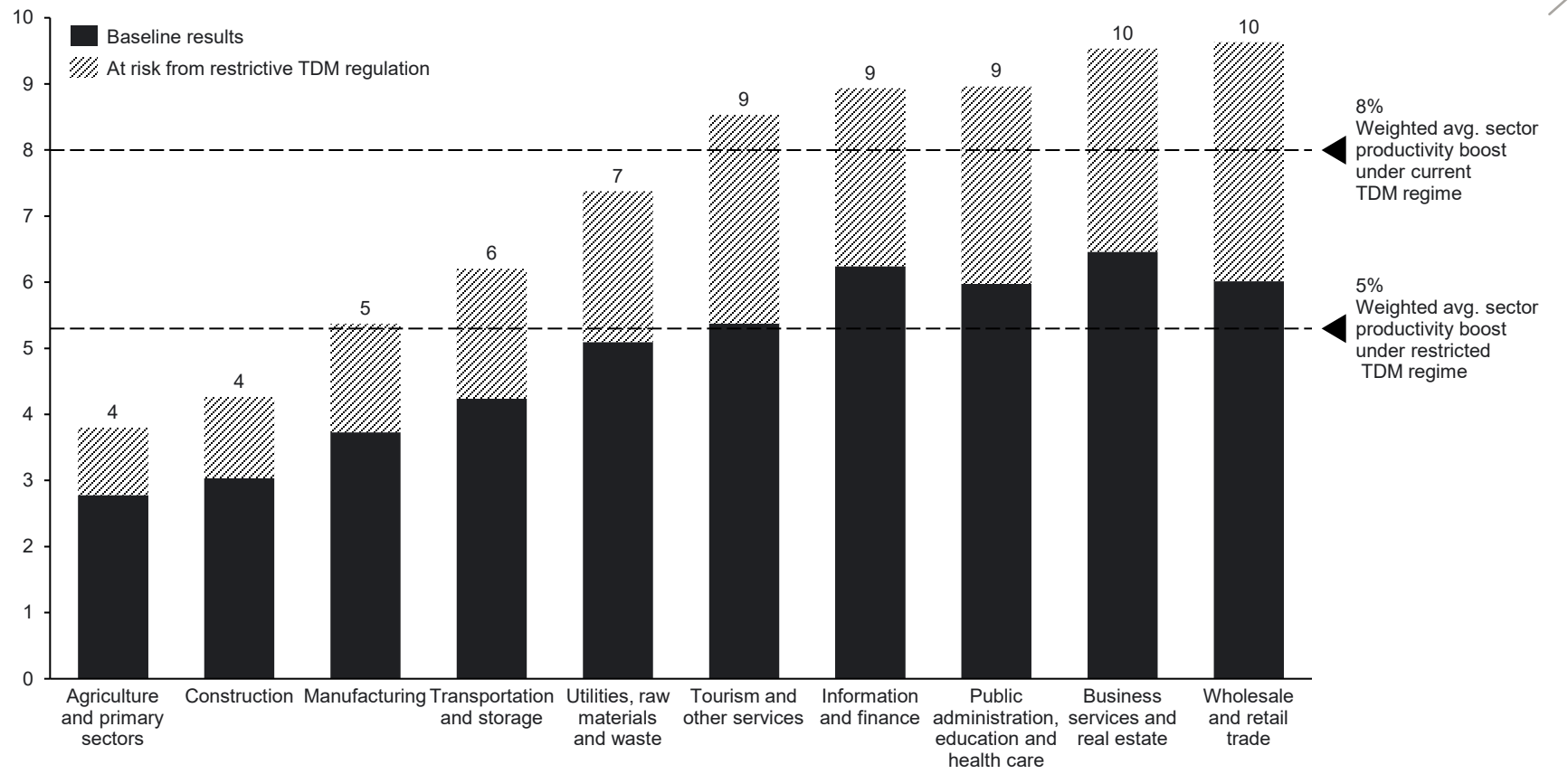
A licensing regime could impose additional transaction costs and redistribute value without creating clear new economic gains, while **disproportionate transparency requirements** and **unworkable opt-outs** would add further friction. Together, these measures could weaken AI models and raise barriers for European developers, putting a further EUR 425 billion at risk.

Source: Implement Economics based on O*NET, Apertus (2025), Briggs and Kodnani (2023).

Knowledge-intensive industries would be most affected by commercial TDM restrictions

Productivity boost from adoption of generative AI

% of sector baseline value added



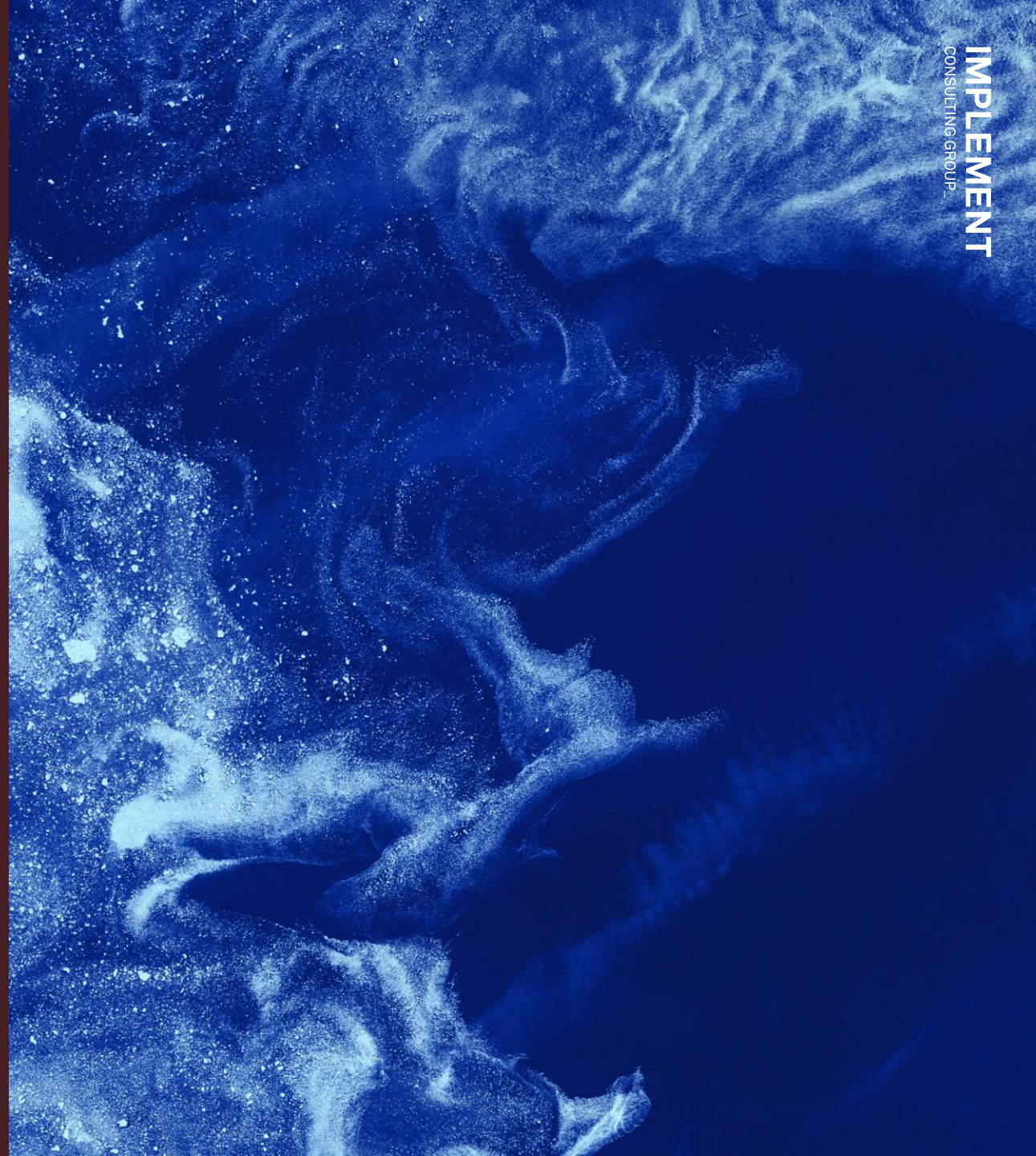
- Restrictions to the current commercial TDM regime will disproportionately impact AI-exposed industries.
- The sectoral variance is driven by task composition. Industries relying heavily on complex reasoning and continuous data access for automation are highly vulnerable to AI bottlenecks.
- The wholesale and retail trade sector is among the most affected industries, materially reducing the potential productivity boost from 10% to just 6%.
- Other knowledge-intensive sectors face similar constraints. Business services, finance, and public administration could all see their baseline productivity gains almost halved under stricter TDM laws.
- This would impede the strategic objectives under the EU's Apply AI Strategy and AI in Science Strategy.
- Overall, enforcing restrictive commercial TDM policies is projected to drag the economy-wide average productivity boost down from 8% to 5%.
- Workable policies remain critical to preventing structural disadvantages in Europe's most valuable service sectors and their global competitiveness.

Source: Implement Economics based on O*NET, Apertus (2025), and Eurostat.

The way forward

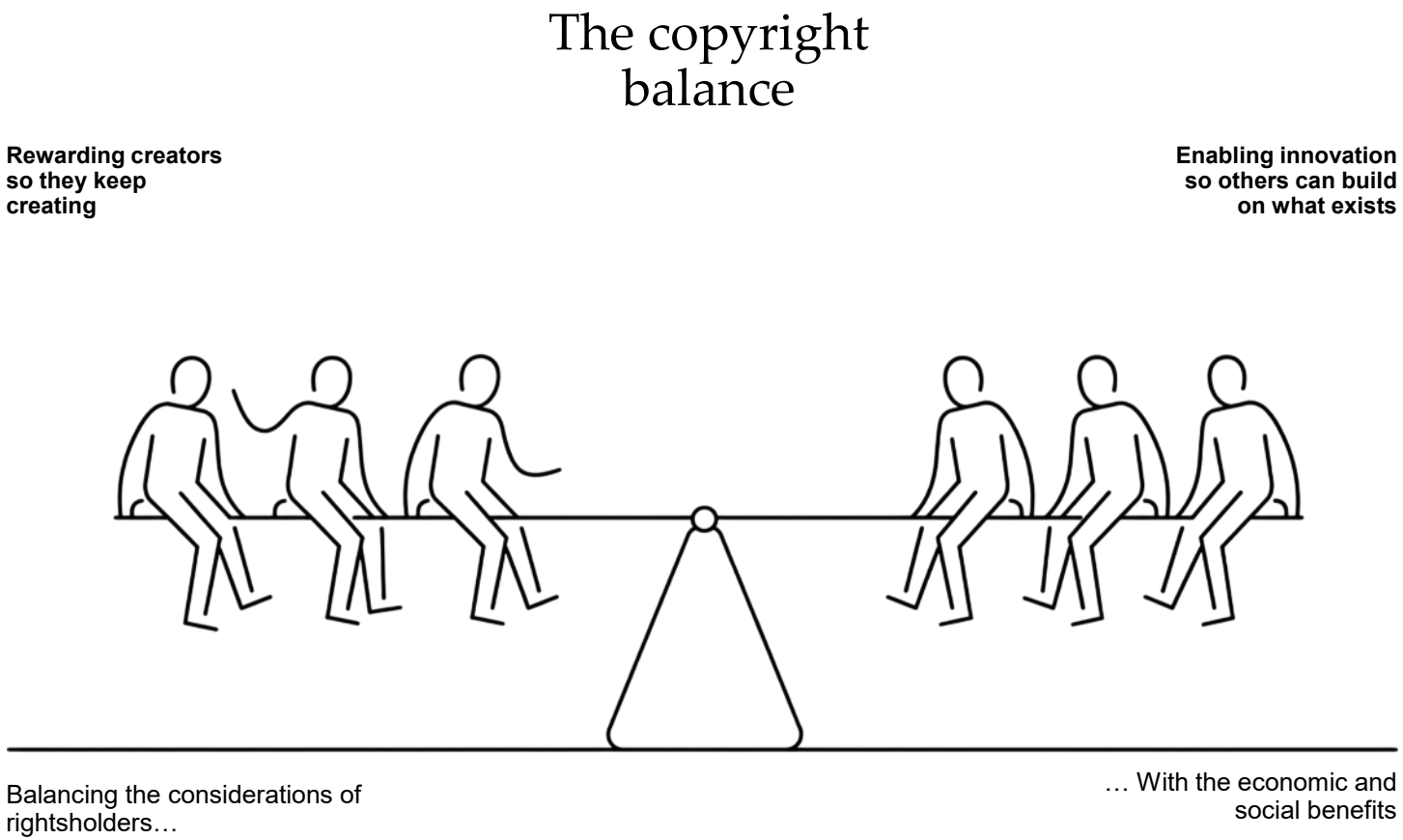
Overly restrictive regulation on copyright and AI risks diminishing Europe's AI potential, so policymakers must consider both the incentive to create and the broader impact on innovation.

> *How wonderful that we have met with a paradox. Now we have some hope of making progress.*
> **Niels Bohr**



IMPLEMENT CONSULTING GROUP

Strict copyright laws could hold back Europe's AI growth, and policymakers need to find a balance between preserving the incentive to create and encouraging new technology



- Europe's AI opportunity is large, but EUR 600 billion in annual value is at risk if the commercial TDM regime becomes more restrictive. The policy question is therefore not whether creators should be protected, but how that protection is designed.
- The EU copyright regime is already balancing the two inherent goals of any copyright regime: Rewarding creators so they keep creating *and* enabling innovation so others can build on what exists.
- A workable regime should therefore both preserve commercial TDM and leave room for voluntary, commercial partnerships where they can create value for rightsholders without undermining scalable AI development.
- The current Article 4 TDM regime allows rightsholders to opt-out through technical protocols, which are machine-readable and scalable, such as the widely used and understood *robots.txt* protocol.[1]
- New frictions would carry real economic costs. Regulatory uncertainty is estimated to put EUR 175 billion at risk through slower adoption, while disruptive initiatives, such as a licensing regime, unworkable opt-outs or disproportionate transparency requirements could put EUR 425 billion at risk by weakening models and raising barriers for European developers.
- This carefully considered balance should remain workable in practice.

Note: 1) The *robots.txt* protocol is used by around 84% of sites. One can think of *robots.txt* as a visitors' centre for each site. It is a file that sits at the root of an origin and allows site owners to implement the Robots Exclusion Protocol. It is designed to instruct bots about which sites it can or cannot crawl. *Robots.txt* file has been used since 1994 to control how a site is crawled, and it became formally standardised with the Internet Engineering Task Force in September 2022. See Web Almanac for details.

To capture the full AI potential, the EU must maintain its commercial TDM, keep opt-out mechanisms workable, and give room for commercial partnerships to grow

The current regime supports large-scale AI development in Europe via the possibility for commercial TDM for training of AI models and applications. The existing regime strikes a crucial balance. It empowers developers to build AI models and applications while cutting the regulatory friction that could prohibit or stall deployment and subsequent adoption of AI tools across Europe. Ultimately, the current TDM approach keeps technological innovation in Europe, allowing AI to develop to fit the needs of EU businesses and users and reflecting European values and cultural norms. Restricting TDM would risk pushing European AI innovation overseas and losing that cultural representativity. Finally, the current regime has paved the way for *robots.txt* as the standardised way for rightsholders to manage their content.

On this background, EU policymakers should…

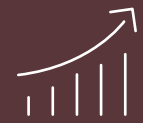
Maintain the current commercial TDM exception

- Preserve the current commercial TDM exception as the present regime already strikes a workable balance between creator protection and innovation.
- Ensure that any transparency requirements are proportionate, operationally feasible, and compatible with AI development at scale.
- Avoid adding new legal frictions that would delay AI adoption, deployment and development and reduce Europe's economic gains, putting up to EUR 600 billion at risk through added uncertainty, transaction costs, and deployment barriers.

Keep opt-out mechanisms workable

- Keep a scalable, machine-readable, and globally recognised opt-out standard across the EU.
- Such a standard already exists, namely the universally available protocol *robots.txt* which is used and respected for rightsholders to opt-out from TDM and prevent the use of their content to train AI models.
- A consensus-driven approach via global bodies will ensure technical functionality, prevent market fragmentation and ensure the mechanism is supported by/ agreed to by both rightsholders and AI developers, including smaller firms, with low compliance costs and legal certainty.
- The use of the existing global standard, *robots.txt*, will provide legal certainty for European AI developers and avoid a situation with fragmented, immature or unproven alternative opt-out mechanisms.

Give room for commercial partnerships to develop further



- Do not enact mandatory or centrally administered licensing models for AI training. This would significantly raise transaction costs, fragment access to data, and weaken AI capability without clear evidence of improved incentives for content creation. It would also require additional policy decision about which content belongs to whom.
- Early signs show that voluntary, commercial partnerships and data-sharing collaborations are already developing to support specialised applications.
- Continuation of commercial partnerships and data-sharing collaborations can also strengthen the incentive for new content creation, while preserving the workable TDM regime that supports AI development, diffusion, and innovation in Europe.

Annex

IMPLEMENT
CONSULTING GROUP_

Literature | Data access and model capabilities

The role of training data

Large language models depend on text and data mining because model capability is shaped not only by compute and architecture, but also by the scale, diversity, and quality of the data used throughout training. In practice, restrictions on training data affect model performance through two main channels. First, they reduce the volume of data available for pretraining, which limits the breadth of knowledge and representations the model can acquire. Second, they constrain the availability of high-value supervised fine-tuning data, which is critical for instruction-following, multi-step reasoning, and other behaviours that make models useful in real-world settings.

Pretraining and data volume

At the pretraining stage, the central issue is scale. A growing body of evidence shows that language models continue to improve when trained on more data, even well beyond earlier “compute-optimal” assumptions. Sardana et al. (2024) find that model quality continues to rise at token-to-parameter ratios up to 10,000, while Gadre et al. (2024) confirm that this pattern holds reliably across a large set of models. Consistent with this, the average tokens-per-parameter ratio for open-weight models rose sharply from 2022 to 2025, while frontier training compute also continued to grow rapidly (Somala and Edelman, 2025; Epoch AI, 2024).

Data volumes used by frontier models

This matters because the frontier is now built on very large and increasingly diverse training corpora. The technical evidence reviewed here shows training runs of around 15 trillion tokens for LlaMA 3, 15.6 trillion tokens for LlaMA 3.1 405B, and 36 trillion tokens for Qwen 3, with GPT-5 estimated at 30 trillion tokens or more (Grattafiori et al., 2024; Yang et al., 2025; You, 2025). These training budgets are far above what was typical only a few years ago, and they reflect the fact that broader and larger datasets remain a direct route to stronger model performance.

Availability of openly licensed data

By comparison, the pool of clearly permissive data is much smaller. Large open or permissively licensed corpora such as Common Corpus and Common Pile are each around 2 trillion tokens, while Wikipedia is around 20 billion tokens, far below the scale used in current frontier training (Langlais et al., 2025; Kandpal et al., 2025). Broader web-scale datasets such as CommonCrawl, FineWeb, RedPajama-v2, and DCLM-Baseline are far larger, but much of that material would not remain available under a more restrictive permission-based regime (Penedo et al., 2024; Li et al., 2024; Together AI, 2023). As a result, restricting training to clearly open or permissive material would reduce the effective pretraining pool well below what leading developers currently use.

Filtering levels and benchmark outcomes

Benchmark comparisons suggest that modest opt-out filtering under the current regime has limited pretraining cost, because the share of removed data is still relatively small. Hernández-Cano et al. (2025) estimate that retroactive enforcement of existing opt-outs removes roughly 8 percent of English-language tokens, and an 8 billion parameter compliant model trained with this level of filtering remains close to a similarly sized benchmark trained on unrestricted data. By contrast, a model trained only on around 2 trillion open-license tokens scores materially lower on MMLU and other capability benchmarks than comparable models trained on much larger corpora. Kandpal et al. (2025) also find weaker results on benchmarks such as HellaSwag and PIQA, which is consistent with a narrower and less representative training distribution.

Constraints at the post-training stage

The more important constraint, however, may arise after pretraining. Pretraining gives a model broad language knowledge, but supervised fine-tuning and related post-training stages are what make that knowledge usable. These stages teach the model to follow instructions, reason through problems step by step, and produce answers in forms that are helpful to end users. Evidence from Apertus shows that when license filtering is applied to fine-tuning data, performance on MMLU CoT-strict falls from 0.513 to 0.253, a drop of ~50%, even though the underlying base model remains the same (Hernández-Cano et al., 2025).

Effects on reasoning

This suggests that the largest capability loss under restrictive data access may not come from weaker pretraining alone, but from the removal of high-value post-training datasets that teach the model how to reason and respond effectively. The capabilities most affected are instruction-following, chain-of-thought reasoning, and other forms of structured problem-solving, rather than basic memorised knowledge alone. That distinction matters because these are precisely the capabilities that underpin useful assistants, enterprise tools, and more advanced forms of automation.

Comparison with frontier models

For context, Wang et al. (2024) show that GPT-4o scores 0.726 on MMLU-Pro, a harder reasoning benchmark than standard MMLU. That gap is important because it highlights how far license-filtered open models remain from the capability level of leading systems on the types of tasks that matter for professional applications. Restricting access to post-training data therefore does not just make models marginally worse. It weakens the specific layer of capability that determines whether a model can perform reliably on demanding cognitive tasks.

From benchmarks to professional tasks

The relevance for deployment becomes even clearer when benchmark performance is compared with economically meaningful tasks. Patwardhan et al. (2025) introduce GDPval, a benchmark that evaluates models against human professional deliverables across 44 occupations covering around $3 trillion in annual US GDP. On this benchmark, GPT-4o achieves a 12.5 percent win rate against human experts, GPT-5 reaches 39.0 percent, and GPT-5.4 is reported at 83.0 percent. These results suggest that economically valuable professional performance emerges only at much higher capability levels than those reached by today’s license-constrained open models.

From capability to economic value

If a model with substantially stronger reasoning performance than compliant open alternatives is still only modestly competitive on professional tasks, then models suffering large reasoning penalties from restrictive data access are unlikely to reach the threshold needed for widespread high-value deployment. The exact mapping from benchmark scores to business value is never perfect, but the technical direction is clear. Restrictions on training data do not simply reduce the quantity of material available to developers. They reduce the ability of models to acquire and apply the reasoning capabilities that matter most for real-world use.

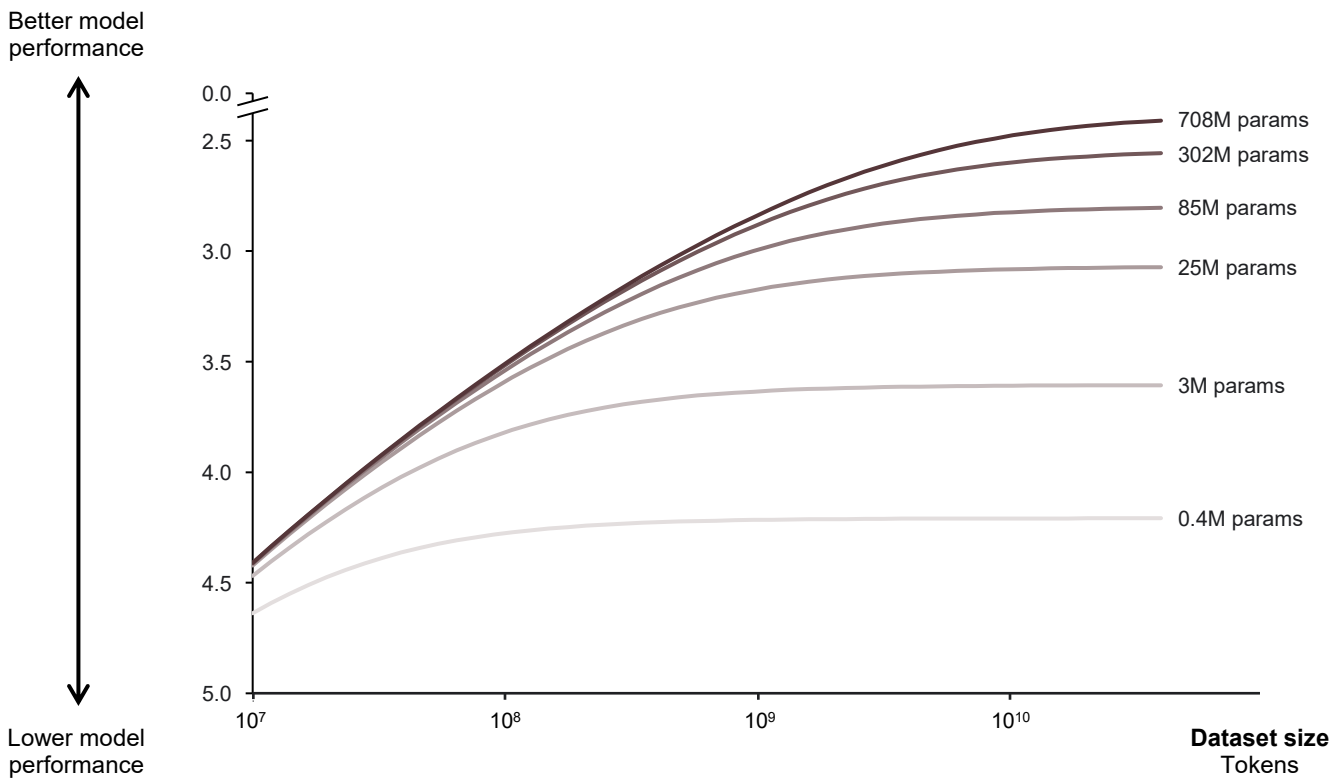
Overall picture

Taken together, the technical evidence points in one direction. Restrictions on training data reduce the volume and diversity of text available for pretraining, but they also remove many of the supervised examples that are most important for making models useful in practice. The result is not simply a smaller dataset. It is a structural reduction in model capability, especially in the reasoning and instruction-following behaviours that matter for business use, research applications, and advanced professional tools. That is the technical channel through which restrictions on text and data mining can translate into weaker AI performance and lower economic value.

Sources: Implement Economics based on Hernández-Cano et al. (2025); Sardana et al. (2024); Gadre et al. (2024); Grattafiori et al. (2024); Penedo et al. (2024); Li et al. (2024); Langlais et al. (2025); Kandpal et al. (2025); Wang et al. (2024); Patwardhan et al. (2025); Somala and Edelman (2025); Epoch AI (2024); Yang et al. (2025).

Size matters – bigger datasets and more parameters improve AI model performance

Dataset size and model performance
Nats/token (test loss unit)



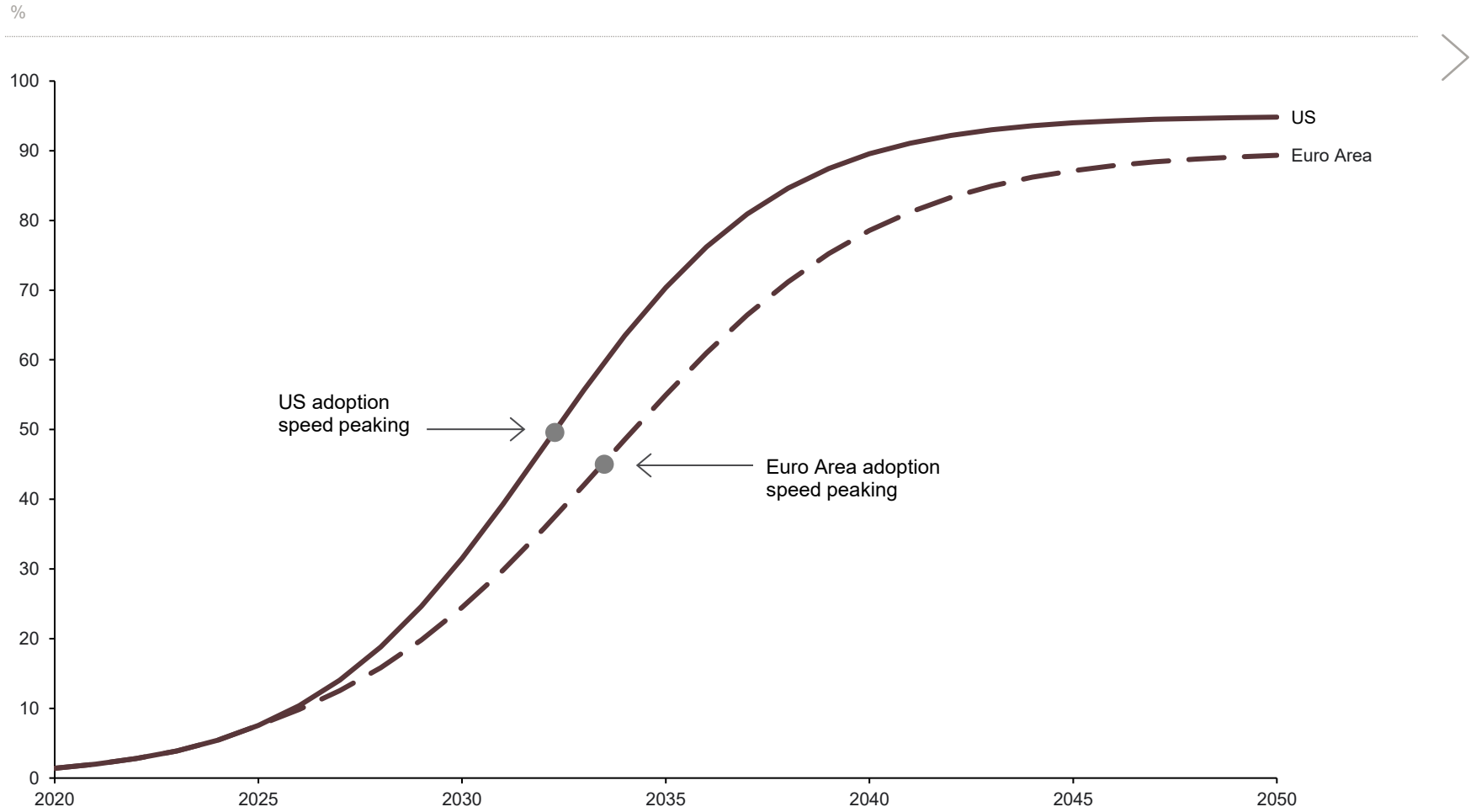
Increasing the size of the training dataset while also expanding the size of the model produces large increases in model performance measured by loss.

- A landmark 2020 study by Kaplan et al. established that language model performance scales as a precise power law with dataset size, meaning each incremental increase in training data yields a measurable, predictable improvement in how well the model handles language tasks.
- Performance in this context is measured by "test loss," expressed in nats per token, which captures how “surprised” the model is by new text it has not seen before. Lower loss translates directly into better real-world capabilities: more accurate answers, more coherent reasoning, and stronger performance on downstream tasks like translation, summarisation, and question answering.
- The Kaplan study found that the scaling relationship between data and performance holds across more than two orders of magnitude in dataset size, from roughly 22 million to 23 billion tokens, with no sign of the trend flattening at the upper end.
- These findings have become foundational to how the AI industry allocates resources, and they underscore a straightforward policy implication: any regime that meaningfully reduces the volume or diversity of data available for training will produce measurably less capable AI systems.

Note: Charts adapted from Kaplan, J. et al., "Scaling Laws for Neural Language Models," OpenAI, January 2020 (arXiv:2001.08361). Test loss is measured in nats per token, a standard metric for language model performance where lower values indicate better performance. X-axes show dataset size (billions of tokens), compute (PF-days), and model parameters (non-embedding) on logarithmic scales.
Source: Implement Economics based on Kaplan et al. (2020)

Full adoption of generative AI is expected to take place over 20-25 years, with the EU expected to achieve adoption 2-3 years after the US

Adoption of generative AI



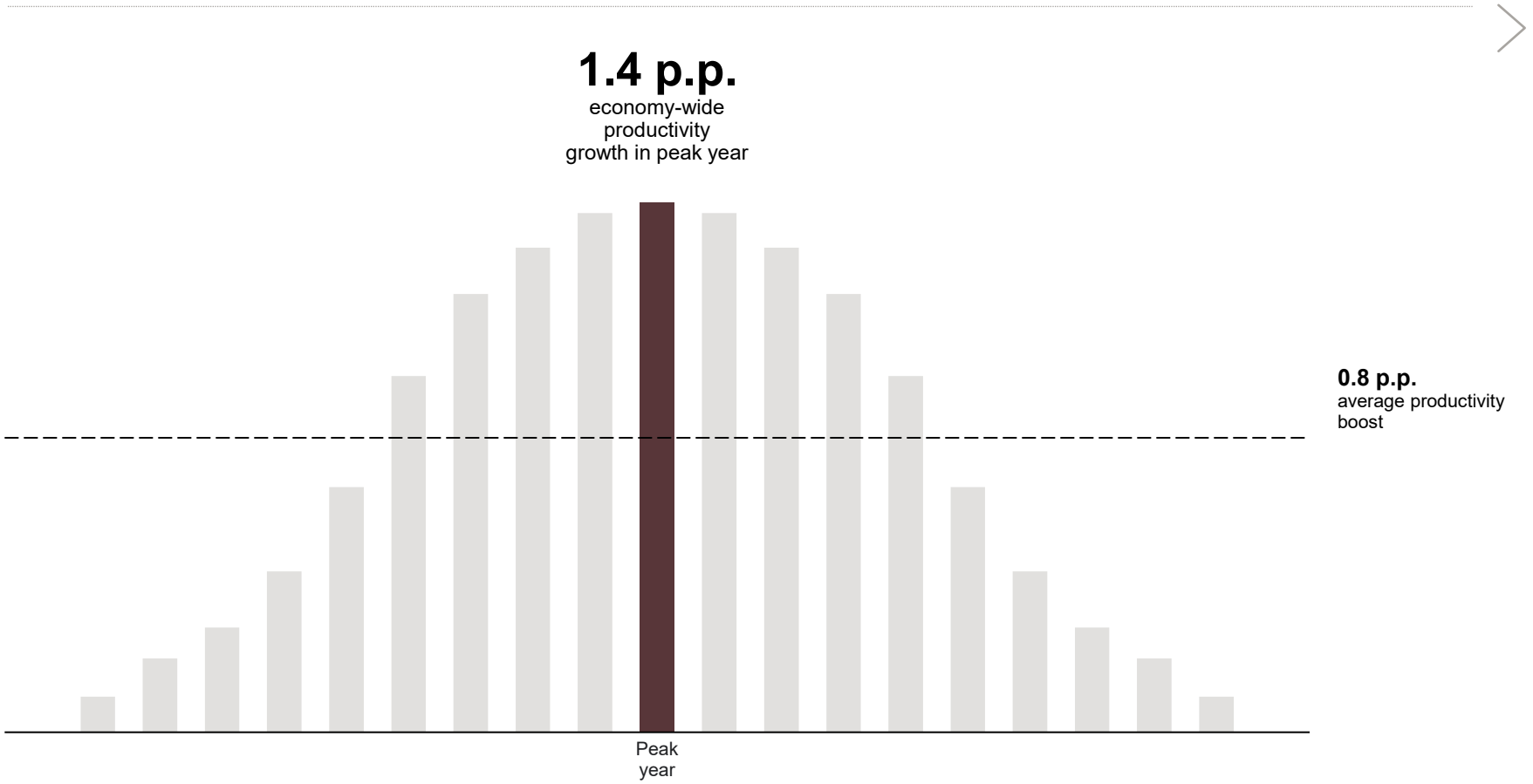
- Compared to other historical technological innovations, generative AI is more powerful, more user-friendly and easier to adopt.
- Widespread adoption (50% adoption) is expected to be reached in the mid-2030s. However, full adoption is expected to take place over a longer period, namely 20-25 years.
- The speed of adoption will be at its highest in the middle of the adoption period in about ten years from now, and this is when the impact on the economy will have its peak year.
- In line with other historical technological revolutions (e.g. electricity and the steam engine), new technology is expected to spread first in the country leading its development – in this case, the US.
- Developed markets in the EU are expected to follow the US adoption rate with a 2-3-year lag, with an expected peak in the mid-2030s.
- In line with Briggs and Kodnani (2023a), this report assumes that the Euro Area follows this trajectory.
- Emerging markets, including Central and Eastern Europe, are expected to adopt the new technology at a slower rate, expectedly peaking in the second half of the 2030s.

Source: Implement Economics based on Briggs and Kodnani (2023).

The annual productivity boost from generative AI will peak in the middle of the adoption period when annual increase in adoption is at its highest

Productivity boost in the EU during the adoption period

Percentage point productivity growth p.a. during generative AI adoption period



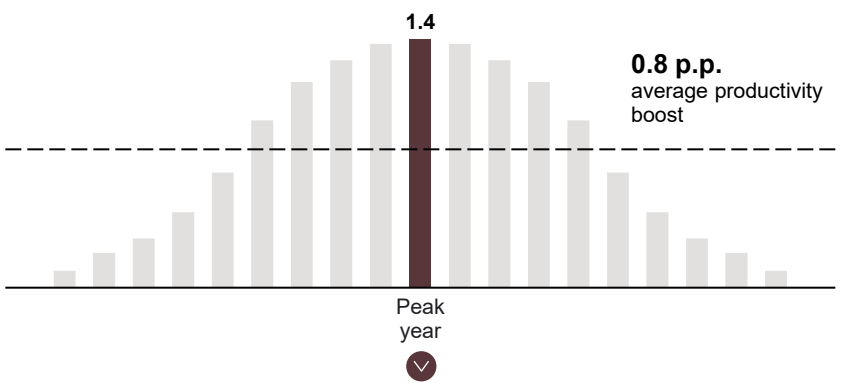
- At its peak, the productivity effect of generative AI in the EU is estimated to be equivalent to 1.4 percentage points annually.
- In the years preceding and following the peak year, the annual impact will be smaller, and very small in the early years.
- Over the full adoption period, the average annual gross productivity potential of generative AI is estimated at around 0.8 percentage points.

Note: The estimate assumes widespread adoption of generative AI over a ten-year period. There is much uncertainty around the capability and adoption timeline of generative AI. The size of the productivity boost depends on the difficulty level of tasks that generative AI will be able to complete and the number of jobs it can automate. GDP is in 2022 levels. The average number of work activities that potentially can be performed by generative AI across all types of tasks for both complemented and highly exposed workers corresponds to 20-25%. Our estimate is the isolated potential of generative AI. The estimated boost from generative AI may not be fully additive to GDP trends, as the GDP forecast already assumes a growth contribution from new technologies and generative AI may substitute some of that. Also, the boost from generative AI may be partially offset by an underlying growth slowdown.
Source: Implement Economics based on Eurostat, O*NET, Briggs and Kodnani (2023).

Generative AI is potentially powerful enough to boost GDP growth in the coming decade with a 0.8 p.p. gross annual potential and a 0.35 p.p. net productivity boost after ICT offsets

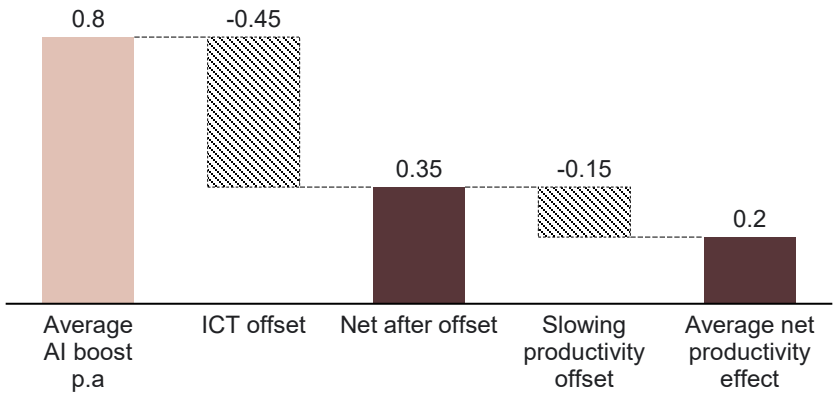
Productivity boost in the EU during the adoption period

Percentage point productivity growth p.a. during generative AI adoption period



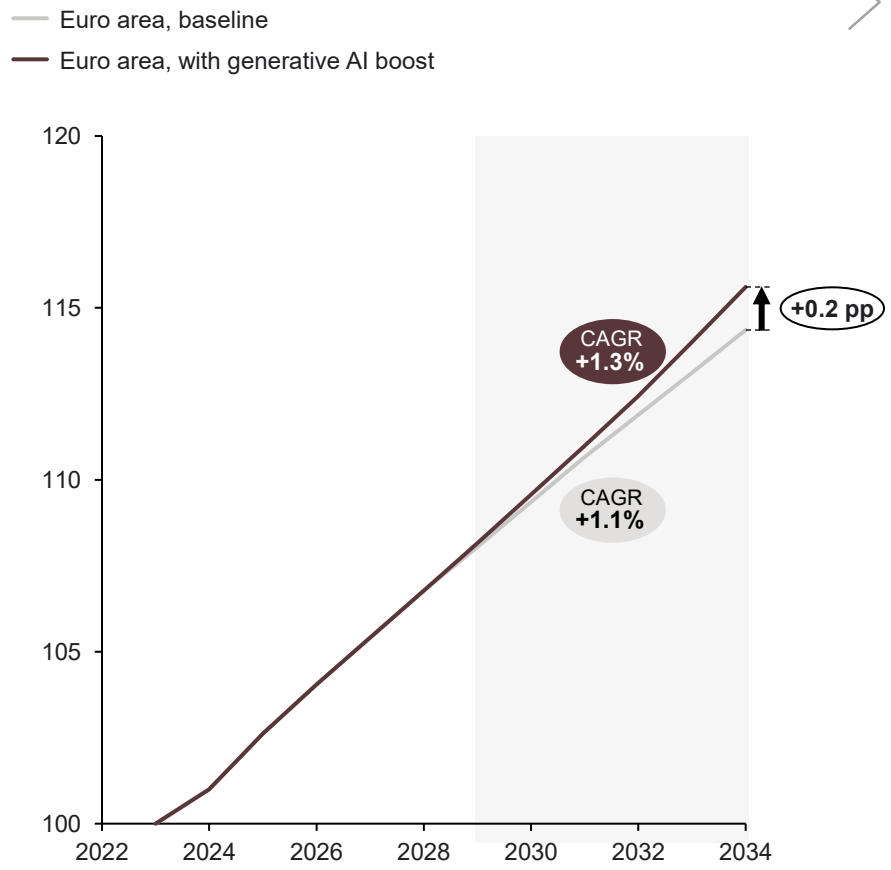
Net productivity boost from generative AI effect after offsets

Percentage point productivity growth p.a.



Euro area GDP, baseline and with generative AI boost

Indexed to 2023 GDP



- When assessing the macroeconomic impact of generative AI, the gross potential of 0.8% per year on average should be considered against the impacts from other ICT that it will replace.
- In the macro-estimates, subtracting the ICT offset leaves a net average productivity boost from generative AI of around 0.35% per year.
- This net effect is comparable to other net impacts estimated by Bergeaud (2024), estimating a net impact of around 0.3% per year when assuming that only around 40% of the technical gross potential can be achieved cost-effectively.
- Reviewing the potential of generative AI, Goldman Sachs raised its long-term GDP growth forecast for the Euro Area by +0.1 percentage points per year in 2028-2030, +0.2 percentage points in 2031-2033 and +0.3 percentage points in 2034, which is the expected peak year.
- On average, this raises the long-term average GDP growth rates in the euro area from 1.1% per year to 1.3% per year when including the updated estimates.
- The boost from generative AI is significant enough to outweigh the otherwise slowing productivity trend (-0.15 percentage points p.a.) – even considering its offsetting effect on the contribution from other ICT.

Note: The offsetting effect refers to generative AI displacing investment in other ICT technologies. However, it still results in a net positive impact on productivity and economic growth.
Source: Implement Economics based on Briggs and Kodnani (2023), Bergeaud (2024), European Central Bank, European Investment Bank (2024) and European Commission (2023).

Studies predict AI will boost productivity, but by how much is uncertain

	Gross effects			Net effects					
Study	**Cumulative gross contribution to GDP** % of GDP	**Gross productivity boost in peak year** p.p. per year	**Gross average productivity boost** p.p. per year	**Cumulative net contribution to GDP** % of GDP	**Net average productivity boost** p.p. per year	**Measure**	**Time horizon for impact**	**Geography**	**Key difference in assumption**
Implement Consulting Group (based on GS model)	8%	**1.4**	0.8	n.a.	0.35	Labour productivity	10 years	EU/euro area	Widespread adoption in 10 years 40-45% of technical potential after ICT offset
Bergeaud (2024)	*7.2%*	n.a.	*0.7*	**2.9%**	0.29 (0.13-0.45)	TFP	10 years	euro area	40% of technical potential can be achieved
Acemoglu (2024)	*2.6%*	n.a.	*0.26*	0.7%	**0.07**	TFP	10 years	US	23% of technical potential can be achieved
IMF (2025)	*6.1%*	n.a.	*1.2*	**1.1%**	*0.22*	TFP	5 years	Europe (31 countries)	18% of technical potential can be achieved
Aghion & Bunel (2024)	*6.8-13.0%*	n.a.	0.68-1.3	*3.4-6.5%*	*0.34-0.65*	TFP & Labour Productivity	10 years	Advanced Economies	50% of technical potential can be achieved

Bold: headline number in publication // Normal: additional metric reported // *Italics*: calculated for comparison (not reported)

Note: Quantifications of the potential of generative AI vary across methodologies, assumptions and economies. *Econometric* studies can quantify existing generative AI productivity boosts but are still dependent on econometric design and a limited empirical database. *Potential* studies, such as Implement's and those shown in the table, quantify the economic benefits from AI from an acknowledged theoretical framework.
Source: Implement Economics based on Acemoglu (2024), IMF (2025), Bergeaud (2024), Briggs and Kodnani (2023), Aghion & Bunel (2024), and Implement Consulting Group (2024).

Modelling the impact of commercial TDM restrictions on the economic potential of AI

Overview of the methodological approach to estimate the economic potential from GenAI and the value at risk from TDM restrictions

1

Automation potential of work activities: The exposure to generative AI is calculated by breaking down the automation potential of unique task descriptions and their associated general work activity in the occupational task database O*NET. In line with Briggs and Kodnani (2023), the methodology assumes that 13 of 41 overall work activities (e.g. getting information, performing administrative activities etc.) can potentially be automated by generative AI, and in the base scenario we assume that tasks with a difficulty up to level 4 on the O*NET-defined scale can be automated.

2

Mapping the automation potential of work activities to occupations: First, the 41 work activities for 900 US occupations are mapped using importance-average activities for each occupation, providing an estimate of the share of each occupation's total workload that AI has the potential to automate. Secondly, this number is projected from US to European occupations through the European Commission's crosswalk between ESCO and O*NET and finally compiled into aggregated occupations (using the sub-occupation employment). This results in four shares describing the share of the work activities for each occupation expected to fall into each category: No automation, Low AI exposure, Medium AI exposure and Likely replacement.

3

Quantifying economic potential at risk from TDM restrictions: The economic impact of a restrictive TDM regime is modelled across four channels: Adoption speed, Model capabilities, Innovation with AI and AI value chain.

First, regulatory uncertainty and compliance burdens are likely to result in delays to AI adoption and development in the EU. We conservatively model this as a one-year lag to AI adoption, corresponding to a ~7 percentage point lower level of adoption in the year with peak marginal adoption.

Second, to account for the degradation in complex reasoning capabilities caused by restricted training data (as shown by e.g., Apertus, 2025), the threshold for automatable task difficulty within the O*NET framework is downgraded from level 4 to level 3 for complex reasoning work activities that are exposed to automation by generative AI.

Third, this combined reduction in adoption speed and model capabilities reduces the downstream R&D productivity gains and AI-enabled innovation, which puts the AI innovation potential at risk.

Finally, modelling the economic potential at risk for the AI value chain is based on evidence from Peukert (2025) showing that jurisdictions with flexible copyright regimes successfully commercialise far more AI applications, evidenced by significantly higher rates of patent filings and venture creation.

- The analytical framework builds upon the methodology developed by Briggs and Kodnani (2023) and Implement Consulting Group (2024), maintaining consistency with previous reports assessing the economic potential of generative AI.
- The structural modelling of TDM restrictions is based on findings from recent academic and technical literature, calibrating the macroeconomic friction penalties to observed technological constraints.

Note: Methodology is based on Briggs and Kodnani (2023). There is notable uncertainty regarding both the total size of the AI economic potential and the exact implications of future TDM regulations.
Source: Implement Economics based on the European Commission, Eurostat, O*NET and Briggs and Kodnani (2023).

Disclaimer

This report (the “Report”) has been prepared by Implement Consulting Group (Implement). The purpose of this Report is to assess the economic value of text and data mining for Europe’s competitiveness.

All information in the Report is derived from or estimated by Implement’s analysis using proprietary and publicly available information. Google (“The Company”) has not supplied any company data, nor does it endorse any estimates made in the Report. In addition to the primary market research and publicly available data, Implement’s analysis is based on third-party data provided by the Company. In preparing the Report, Implement has, without independent verification, relied on the accuracy of information made available by the Company. Where information has been obtained from third-party sources and proprietary research, this is clearly referenced in the footnotes. The Report is based on work conducted in March and April 2026. Implement will not make any representation or warranty as to the correctness, accuracy or completeness of the contents of the Report or as to the sufficiency and/or suitability thereof for the Company’s or the reader’s purposes, nor does Implement assume any liability to the Company, the reader or any other legal entities for any losses or damages resulting from the use of any part of the information in the Report. The information contained herein is subject to change, completion or amendment without notice. In furnishing the Report, Implement undertakes no obligation to provide the Company with access to any additional information. The Report is an independent report funded by the Computer & Communications Industry Association (CCIA Europe). The opinions offered herein are purely those of the authors and do not necessarily represent the views of CCIA Europe.