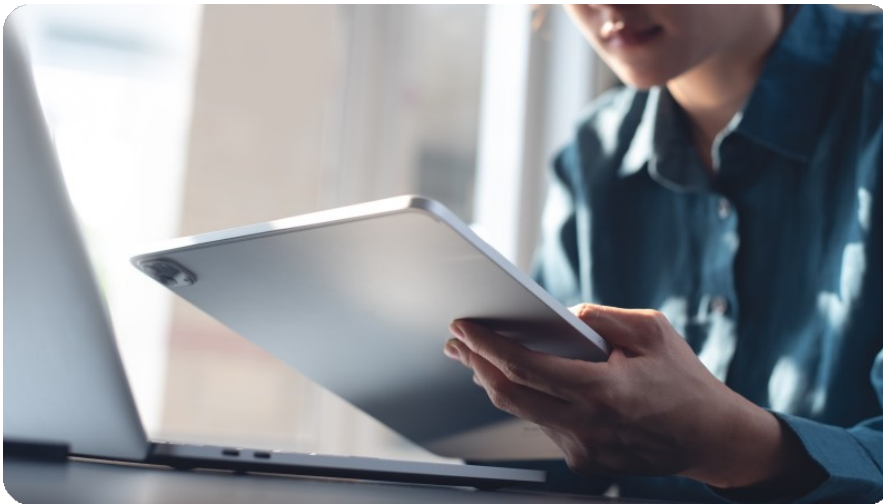


Article | 1 November 2024

TMT

AI Monthly: Soaring costs of GenAI challenge short-term profitability

New data suggests that AI is driving innovation and discovery beyond the Information and Communications Technology (ICT) sector. This is a promising sign, but there are growing concerns about the short-term returns on AI investments



Complementary innovation increases

Generative AI is considered a [general-purpose technology](#), which is a technology that has a profound impact on the economy because it is pervasive, shows rapid improvement, and has the potential to drive complementary innovation across various economic sectors.

Recent research by Damioli et al. (2024) provides evidence for GenAI enabling complementary innovation. Between 2000 and 2016, global AI patenting accelerated significantly and became increasingly pervasive, with a notable shift away from the ICT sector to other areas of the economy.

Interestingly, this surge in patent activity was primarily led by relatively young and smaller companies, indicating that generative AI deployment fosters increased innovation.

Investment in LLMs still booming

The positive news about innovation is tempered by concerns about investment returns. Large

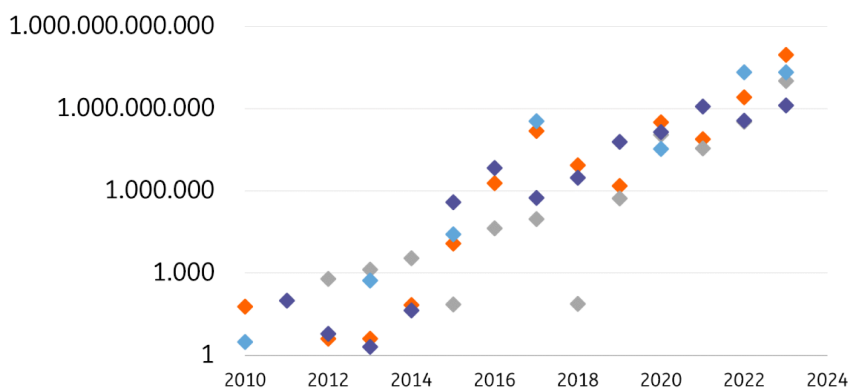
language models (LLMs) have made exponential strides in recent years, but this progress has been accompanied by a corresponding exponential increase in training costs.

In 2017, training a top-of-the-line model cost roughly \$1,000, but by 2024 that cost had risen to around \$200m, despite a rapid decline in computing costs. The driving force behind this surge in training costs is the astonishing growth in computing power required by LLMs.

As model training data expands, the returns to scale in terms of model performance remain constant. This process therefore requires digital infrastructure such as data centres and ultrafast chips. If this trend continues, we could see the first trillion-dollar model before 2030. In addition, the largest investors in AI such as Amazon, Google, Meta, and Microsoft show no signs of slowing down.

Computing power used to train AI models

Computing power in Petaflop (logarithmic scale)



Source: Epoch (2024); Our World in Data

Concerns about investment returns increase

Given these eye-watering investments, it is no wonder that concerns about investment returns are also on the rise. However, estimates about AI's impact on productivity growth vary. Initially, [we predicted](#) 0.1 to 0.5 percentage points of additional productivity growth per year, which is at the lower end of the scale.

But here's the challenge: productivity growth tends to lag behind investment. Yet productivity growth is essential if these large investments are to be recouped. Recently, Martens (2024) highlighted a critical issue: our current investment trajectory is unsustainable unless productivity growth reaches 3% annually. This jump in productivity growth is not likely in the near term, as generative AI implementation requires time and investment from organisations that aim to use the technology.

Investors in AI face a difficult choice

The rapidly rising costs of generative AI are hard to match with near-term profitability. This confronts investors in AI with a conundrum: dial back investments and salvage near-term profitability or continue investing because expected future gains are too significant to miss out on.

Given [the current AI race](#) between the largest tech companies, we think it is unlikely that the largest investors in AI will hold back. Currently, this is feasible given their very profitable (cloud) businesses. Microsoft this week announced that cloud revenue in the second quarter rose 23% year-on-year. However, if current investment trajectories continue, the financial risks taken by these companies become ever-larger. This, in turn, poses an increasing risk to the financial health of these companies and a systemic risk for the tech industry.

Author

Diederik Stadig

Senior Economist, Healthcare & Technology
diederik.stadig@ing.com

Disclaimer

This publication has been prepared by the Economic and Financial Analysis Division of ING Bank N.V. (“ING”) solely for information purposes without regard to any particular user’s investment objectives, financial situation, or means. *ING forms part of ING Group (being for this purpose ING Group N.V. and its subsidiary and affiliated companies)*. The information in the publication is not an investment recommendation and it is not investment, legal or tax advice or an offer or solicitation to purchase or sell any financial instrument. Reasonable care has been taken to ensure that this publication is not untrue or misleading when published, but ING does not represent that it is accurate or complete. ING does not accept any liability for any direct, indirect or consequential loss arising from any use of this publication. Unless otherwise stated, any views, forecasts, or estimates are solely those of the author(s), as of the date of the publication and are subject to change without notice.

The distribution of this publication may be restricted by law or regulation in different jurisdictions and persons into whose possession this publication comes should inform themselves about, and observe, such restrictions.

Copyright and database rights protection exists in this report and it may not be reproduced, distributed or published by any person for any purpose without the prior express consent of ING. All rights are reserved. ING Bank N.V. is authorised by the Dutch Central Bank and supervised by the European Central Bank (ECB), the Dutch Central Bank (DNB) and the Dutch Authority for the Financial Markets (AFM). ING Bank N.V. is incorporated in the Netherlands (Trade Register no. 33031431 Amsterdam). In the United Kingdom this information is approved and/or communicated by ING Bank N.V., London Branch. ING Bank N.V., London Branch is authorised by the Prudential Regulation Authority and is subject to regulation by the Financial Conduct Authority and limited regulation by the Prudential Regulation Authority. ING Bank N.V., London branch is registered in England (Registration number BR000341) at 8-10 Moorgate, London EC2 6DA. For US Investors: Any person wishing to discuss this report or effect transactions in any security discussed herein should contact ING Financial Markets LLC, which is a member of the NYSE, FINRA and SIPC and part of ING, and which has accepted responsibility for the distribution of this report in the United States under applicable requirements.

Additional information is available on request. For more information about ING Group, please visit www.ing.com.