NATIXIS
CORPORATE AND INVESTMENT BANKING

EQUITY DERIVATIVES INSIGHT

I think. Therefore AI am.

Surfing the ChatGPT wave…

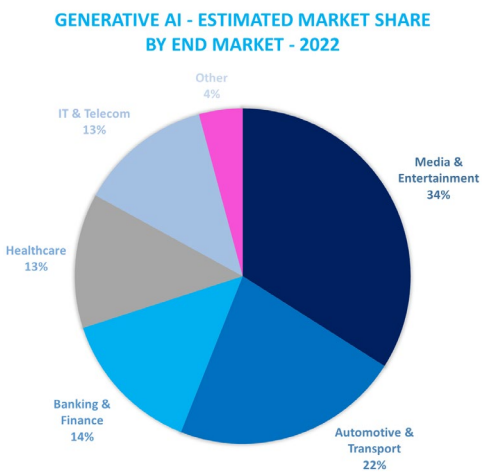
> The most relevant way to build Generative AI equity exposure over the next 2 years, is to invest in the key enabling technologies that are being deployed in the background…

Eric Benoist
Tech & Data Research Specialist
eric.benoist@natixis.com

01 I Generative AI: a game changer for most industry sectors

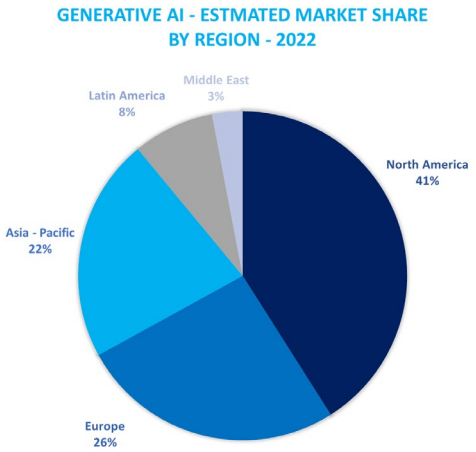
Although generative AI has been a focused field of AI research since 2014 and even has early incarnations dating back to the 1960s such as the ELIZA chatbot of MIT Professor Joseph Weisenbaum, it seems to have mostly appeared on everyone's radar at the end of last year, when Open AI unleashed ChatGPT on the masses, making it the fastest-growing app in history, outpacing the likes of Netflix, TikTok or Instagram. Alongside other image generation services such as DALL-E 2, Stable Diffusion or MidJourney, ChatGPT simply took the world by storm in an almost obscene display of its fantastic abilities.

As the hype shows no sign of abating, what really stands out are the potential **productivity gains** that could result from this fascinating technology. Across many enterprise departments and in a wide range of activity sectors, the possibilities are simply endless. Although it is difficult, at this stage, to quantify their impact, it is likely that these tools will indeed serve to optimize many internal processes. For instance, lower value-added tasks may easily be delegated to machines while employees are given the opportunity to focus on more critical client requests. Elsewhere, content generation or after-sales services may be facilitated or even taken over completely by smart bots, enhancing customer experience and overall product economics. From healthcare and research to gaming, transport, education, IT, finance or legal, **no industry will be left unscathed**.



Sources: Precedence Research, Oddo BHF

As is often the case, the market potential of an emerging technology for which all possible applications are not yet known is difficult to estimate. According to Precedence Research (*Generative AI market, December 2022*), generative AI (software and IT services) will grow **from $11bn in 2022 to $118bn in 2032**, exhibiting an impressive **27% CAGR** over the period. Interestingly, and somewhat consistent with the distribution of global advanced computing infrastructure, much of the activity so far has been emanating from North America (41%), ahead of Europe (26%) and Asia (22%). In our opinion, there is little evidence to date that this hierarchy may change in the future.



Sources: Precedence Research, Oddo BHF

Despite the general excitement, the almost "divine" or "magical" skills of ChatGPT cannot hide the fact that these new breed systems are still subject to **many constraints**. Some are technical in nature, others are more general. All present significant challenges to the successful development of artificial intelligences.

02 I Machine learning limitations…

The use of the plural here is deliberate. There are indeed multiple iterations of AI, each trained differently to excel at a specific task. To understand the scope of what is possible with Generative AI is therefore to understand the models that underlie the technology. Large Language Models (LLMs) are today overwhelmingly based on the **Transformer approach** developed by Google Brain in 2017 and have been designed to solve **sequence transduction problems**, i.e. the transformation of input sequences into output sequences, as in language translation, speech recognition, etc. These are outperforming previous frameworks such as Recurrent Neural Networks (RNN) or Long Short Term Memory (LSTM) thanks to a "**self-attention**" mechanism that allows them to make sense of words and their connections, even in the case of complex semantic dependencies. Processing can also be parallelised for a faster end result. Still, the algorithms involved are far from perfect. The proposed encoding methods do not always capture the full context behind a given 'prompt' and introduce a significant amount of guesswork into the process. As a result, the system sometimes becomes too creative and "**hallucinates**" facts without warning. Depending on the intended use, this may or may not be a problem, but it certainly warrants a **high degree of caution** when dealing with critical business assignments.

Much has also been said about the risks associated with training models on unfiltered data sets. Pre-trained generative transformers have been fed with huge amounts of data often scraped from the internet without any particular consideration for quality or relevance. This unfortunately increases the likelihood of perpetuating **racial, gender and other harmful biases**, as numerous studies have proven in recent years. Awareness raising is good, but remediation is not straightforward, as problems persist even after corrective measures have been taken (through Reinforcement Learning in particular). Data curation does not seem to eradicate the issue either, as researchers themselves may introduce biases in the way they handle training in the first place. Neutrality may simply prove unachievable and sadly, toxic, prejudicial behaviour may never be completely eliminated.

Image: DALL-E 2, a medieval painting of the Wi-Fi not working…

Although transformers are very good at manipulating text strings or performing any activity involving a similar logic (coding, DNA sequence analysis…), **their output cannot be assimilated to quantitative reasoning**. Their probabilistic assessment of a situation and their ability to detect subtle patterns of information are not synonymous with mathematical thinking: in fact, the web is full of amusing ChatGPT responses to mathematical tricks that probably wouldn't fool a primary school pupil and in most cases, when asked to formalise proofs into code, the machine still falls short of human competence. Google's Minerva project is attempting to address this by teaching step-by-step argumentation to LLMs but whether the method can scale to complex problems remains to be seen. Such fundamental weakness must be taken into account by users: generative AI certainly is a powerful tool in specific scenarios but it can in no way replace the human brain in tasks involving advanced Cartesian gymnastics.

At least, not yet. Unsurprisingly, the natural inclination of many data scientists is to believe that the technology is still in its infancy and that these teething problems will be ironed out very quickly. In theory, there is no reason why this should not be the case. In practice, however, the **hardware** dimension must be accounted for.

03 Hardware and data protection challenges…

As transformer architectures become more sophisticated, models are trained on **larger data universes through denser neural links**. GPT-3 (June 2020) was built with 175 billion parameters, Microsoft/NVIDIA's Megatron-Turing NLG (October 2021) had 530 billion and Google's PaLM (April 2022) had 540 billion. Open AI has not yet revealed the size of GPT-4 (March 2023), but it is thought to use over a trillion variable points. Google estimates that the final training "FLOP-count" (Floating Point Operations per second) for PaLM exceeded the 2.5*10^24 mark, using a 6,144-chip optimised TPU configuration for over 50 consecutive days. Even if such compute power is achievable within the boundaries of our current technological mastery, it is far from trivial and not within the reach of everyone. In terms of cost, **it inevitably runs into the millions**: PaLM's training is valued at between $9mn and $23mn by industry experts (*Estimating PaLM's training cost, Lennart Heim, April 2022*).

Furthermore, the Self-Attention mechanism mentioned earlier has **memory requirements** that are quadratic with the input sequence length. This makes it difficult to process long input material (typically > 512 tokens) and forces researchers into developing various "sparse attention" alternatives (such as Google's BigBird) that often sacrifice a degree of expressiveness and flexibility.

All in all, transformer AI is a **supercomputing** concept and, as such, can only realistically be available in the **cloud** through a set of **APIs**. Assuming millions of users start tapping into these systems, as they do with their favourite search engine, the target infrastructure must then be designed for massive, ultra-fast data transfers between internal compute servers and across the network. This requires, amongst other things, significant advances in photonics to enable effective large-scale deployment.

Without a doubt, a cloud-based environment will also cause additional **security concerns**. Sensitive data passing through the platform will need to be securely encrypted and protected from misuse. As things stand, LLMs and the people who run them continue to have unrestricted access to any content provided, in order to train the algorithms and improve the quality of their results.

In the case of sensitive corporate information, it is very difficult for businesses to justify taking such risk and the position of large companies such as JPMorgan, Accenture or Amazon to restrict or ban the use of ChatGPT for business purposes is totally understandable in the absence of a more robust cybersecurity framework.

Generally speaking, data usage policies in the context of generative AI software are a headache for legal and compliance departments around the world. **Copyright issues** begin to emerge, for example when an artist's work is diverted to produce marketing content or an author's ideas are plagiarised. In those cases, it is not clear that standard training practices would fall under the "fair use doctrine", particularly when conducted by for-profit organisations. To complicate matters, each jurisdiction would likely follow their own regulatory path. The same goes with data privacy rules: in Europe, **machine learning is not exempt from obligations under the GDPR** and Personally Identifiable Information (PII) remains out of scope. Are lawsuits inevitable? Probably. Predicting their outcome, however, remains impossible at this stage. Pending greater visibility, take-up is bound to be somewhat more hesitant.

Such considerations necessarily raise the question of the most **viable business approach** for developers of Generative AI solutions. Indeed, it may be prudent to consider building smaller models trained on specialised datasets to solve more targeted puzzles (imagine, for instance, an oncology-focused transformer trained on all existing medical research): those would be easier to bring into compliance – and incidentally – to monetise with professional users. Meanwhile, large-scale generalist AI models may have to bring some services behind a paywall to secure funding for their long term development: subscriptions may be offered in exchange for priority, unlimited access or for various side functionalities and applications. Alternatively, they may form part of a **search engine portal** and start generating ad revenue - although the exact modalities of such combinations would have to be more precisely defined.

04 I A heavy price to pay?

As admirable as research in the name of human progress may be, there is every reason to believe that generative AI will remain in the hands of a highly concentrated technological elite that will derive substantial profit from it. This, of course, is fine as long as the cost to society does not exceed the benefits.

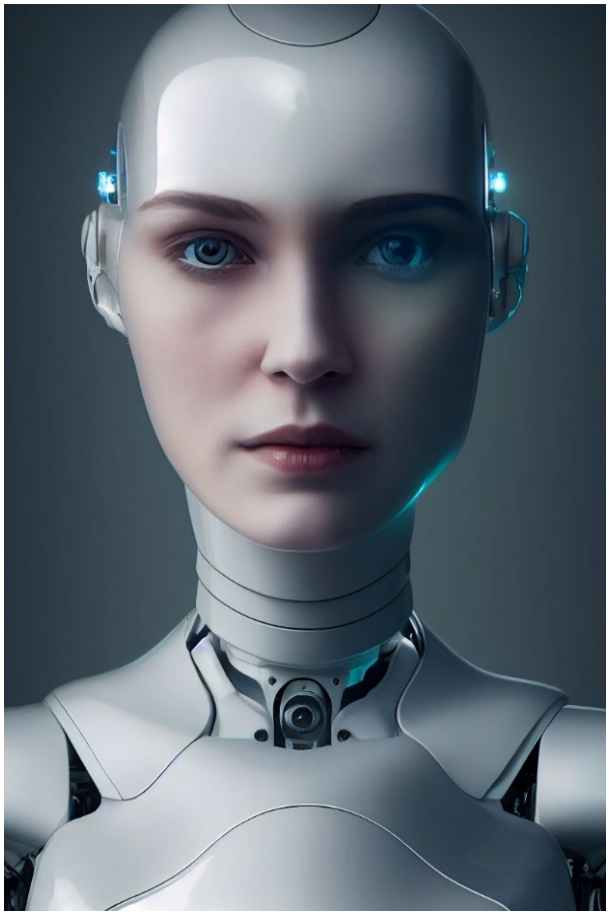


Image: MidJourney, transparent elegant futuristic female android

First of all, this cost has **a strong ecological component** linked to the very energy-intensive nature of the systems in question. Market leaders communicate little about the carbon footprint of their "digital creatures" and observers must therefore rely on the more or less substantiated estimates of various experts in the field. Bloomberg, for example, cites a 2021 study that places the consumption of GPT-3's training at 1.287 GWh, the equivalent of 120 US homes in a year.

According to the same sources, 502 tonnes of CO2 may have been released during the process, equivalent to the annual production of 110 cars. This is all the more disturbing as training is never a one-off event and must be carried out regularly to update a model's knowledge. Worryingly, training may also just be the tip of the iceberg: another analysis published in Towards Data Science (*ChatGPT's electricity consumption, Kasper Groes Albin Ludvingsen, March 2023*) proposes a worst-case scenario ChatGPT electric consumption of 23 GWh based on the number of visits reported in January 2023 (i.e. 590 million)... As the much larger GPT-4 enters the scene, and presumably others are preparing to follow suit, it is worth keeping an eye on the possible excesses of these new instruments.

Finally, there is an obvious **social cost** to implementing these technologies, as **labour will almost certainly be disrupted**. AI proponents often argue that these tools merely improve productivity to allow humans to concentrate on high value-added tasks, and that those who see their work becoming obsolete can simply train in the new techniques to become happy users in their turn. For them, in the end, AI is just another evolution, like the advent of the PC or smartphone before it: society will adapt, as it always has. Society will indeed adapt, but in a non-instantaneous way. This will be a major change of paradigm, a tectonic shift of colossal proportions, one that will span generations and impact low- and high-skilled workers alike. It will certainly not be painless.

This is why it is so important to prepare for it and craft an appropriate **regulatory response** that strikes the right balance between fostering innovation and unleashing uncontrolled disruption on people. Therein lies the greatest challenge of all and the key to a sustainable adoption.

05 | Surfing the wave…

For **equity investors** seeking exposure to this emerging theme, **options are unfortunately limited** at this point. Many of the companies driving the revolution in the field are actually privately held. Similarly, many of the businesses at the forefront of innovation in terms of applications and software powered by generative AI are in the early stages of development and still controlled by venture capitalists.

We could obviously suggest a long list of companies with more or less assertive claims or evolved commercial strategies based on the enthusiastic declarations of Management during their last analyst presentations. This would be, at best, speculative, at worst, outright misleading. These activities are still too underdeveloped to have a significant impact on short-term earnings and the business models mentioned have yet to prove themselves.

In our view, the most relevant way to build such portfolio over the next 2 years, is to **invest in the key enabling technologies being deployed in the background**. These include the **semiconductor** manufacturers leading the race in cutting-edge logic (Nvidia, AMD, Intel…) and memory (Micron, SK Hynix, Samsung…), but also novel microchip-architectures and advanced equipment (BE Semiconductors, ASML, ASMi…). It equally concerns **network infrastructure** and cloud computing specialists without whom massive and instant data transfers wouldn't be possible (Arista, Cisco, Amazon, Alphabet or even Baidu in the Chinese market…). Finally, **software giants** (Microsoft, Salesforce, Adobe, IBM…) or their lesser known - but more focused - competitors (C3.ai, Palantir, Appian…) are all formidable AI pioneers with credible and sizeable ongoing investment plans, making them hard to ignore.

As more generative AI start-ups continue to mature and eventually demonstrate the viability of their platforms, we expect them to enter major stock exchanges and benchmark indices: only then will it become possible to accurately track their performance and publicly participate in their growth.

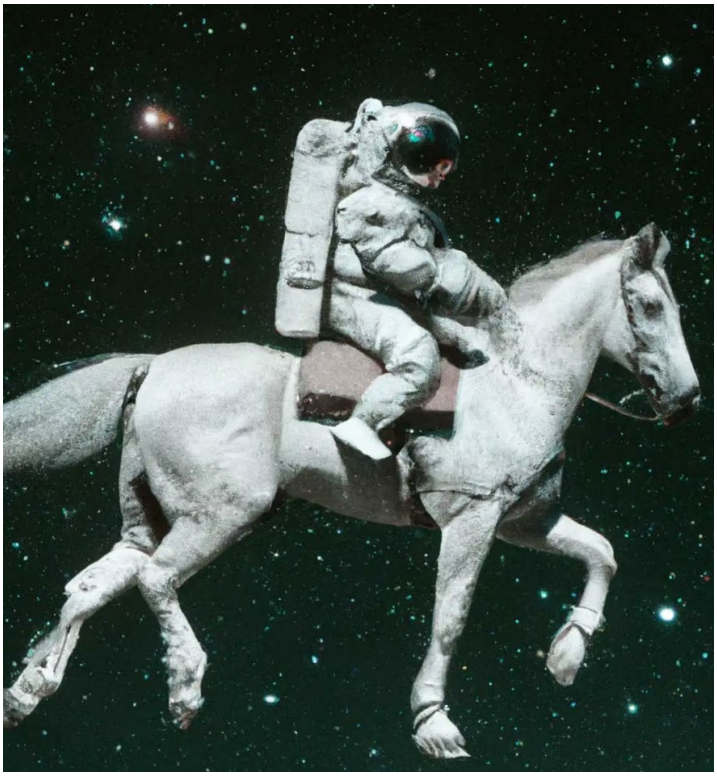


Image: DALL-E 2, photo of an astronaut riding a horse…

DISCLAIMER

LONDON
Cannon Bridge House
25 Dowgate Hill
London EC4R 2YA, UK

PARIS
47, Quai d'Austerlitz
75013 Paris
France

NEW YORK
1251 Avenue of the Americas
New York, NY 10020
USA

SINGAPORE
41st Storey - Singapore Land Tower, 50 Raffles Place
048623 Singapore

MADRID
Paseo de Recoletos 7 y 9,
6ª planta - 28004 Madrid
Spain

FRANKFURT
Im Trutz Frankfurt 55
60322 Frankfurt
Germany

MILAN
Via Borgogna, 8
20122 Milano
Italy

MOSCOW
Tsvetnoi boulevard, bld 2
Entrance C - 4th floor
127051 Moscow, Russia

DUBAI
DIFC Gate Village # 8
Level 5, P.O. Box 506694
Dubai, U.A.E.

HONG KONG
ICC Level 72
1 Austin Road West
Kowloon, Hong Kong

TOKYO
Ark Hills South Tower 8/F
1-4-5 Roppongi
Minato-ku, Tokyo 106-0032