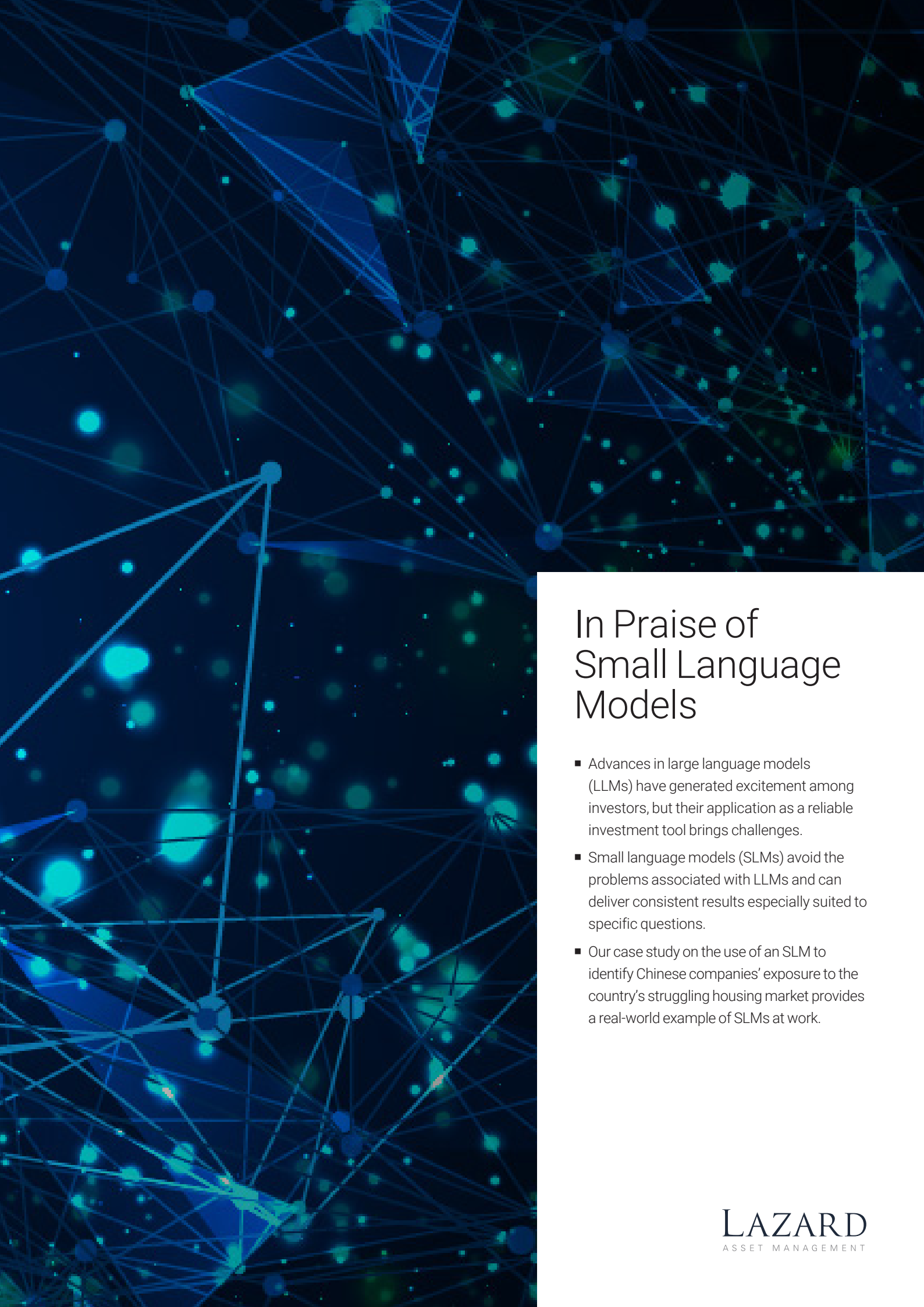


In Praise of Small Language Models

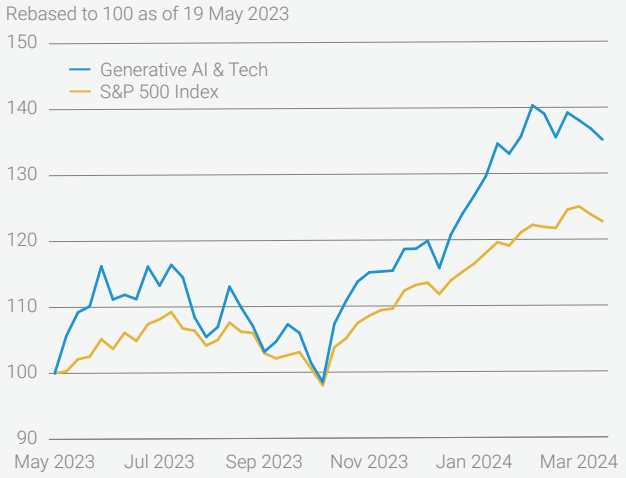
- Advances in large language models (LLMs) have generated excitement among investors, but their application as a reliable investment tool brings challenges.
- Small language models (SLMs) avoid the problems associated with LLMs and can deliver consistent results especially suited to specific questions.
- Our case study on the use of an SLM to identify Chinese companies' exposure to the country's struggling housing market provides a real-world example of SLMs at work.

LAZARD
ASSET MANAGEMENT

Large Language Models are miraculous, but …

A year on from the release of ChatGPT-4, an artificial intelligence (AI) boom has captivated investors and helped propel global stock markets, led by the US. The Roundhill Generative AI & Technology ETF (CHAT), comprising 25-50 companies involved in generative AI and related technologies outperformed a buoyant S&P 500 index by a substantial margin (13%) in the six months to March 2024 (Exhibit 1).

Exhibit 1
Investor Optimism for Generative AI



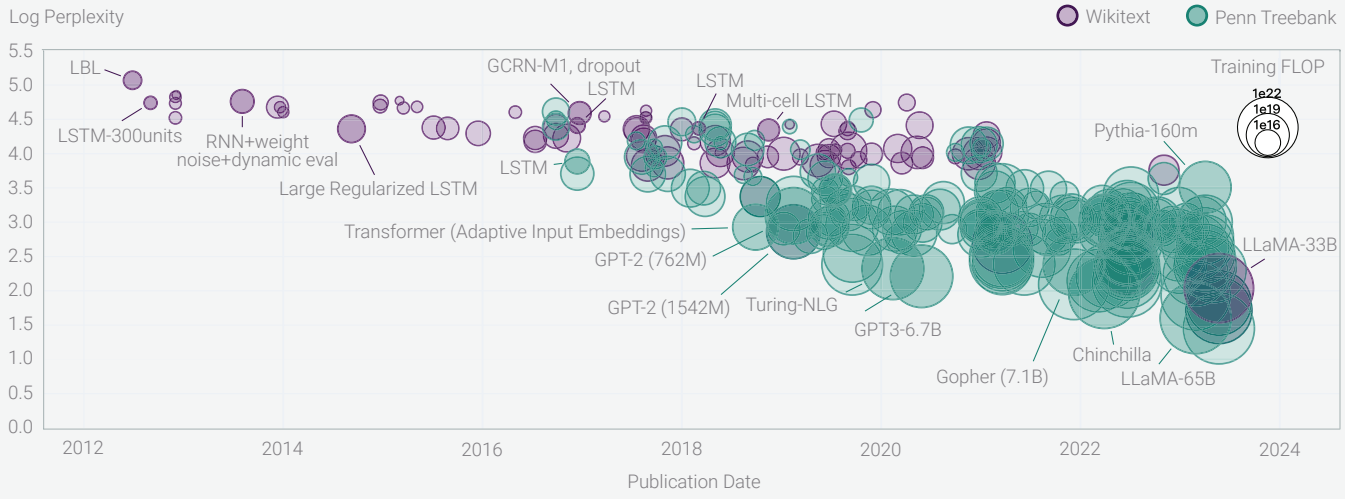
As of 10 April 2024
Source: Bloomberg, CHAT ETF, S&P 500 Index

LLMs, the technology powering this boom, are advancing rapidly. A recent paper highlights that the computational power needed to achieve a given performance threshold is halving every five to 14 months, which is roughly twice the speed of hardware improvements under Moore's Law (Exhibit 2).[1]

And yet two things make it hard to rely on LLMs in the realm of investing: hallucinations and opacity. Hallucination refers to the tendency of LLMs to confidently present made-up information as fact. The problem is part of a broader "alignment" challenge, which aims to make AI reliably follow human intentions. Many researchers think larger, better-quality datasets will fix the issue, but some argue that the inherent generative and probabilistic nature of LLMs makes some hallucinations inevitable. Regardless of which perspective proves correct, the rapid advancement of LLMs suggests that the often embarrassing and costly instances of hallucinations are likely to decrease significantly over time.[2,3]

The second challenge, opacity, refers to the fact that the models operate as "black boxes." With billions (sometimes trillions) of parameters, their enormous scale makes it nearly impossible to fully comprehend their internal logic. And because of the labyrinthine network of interconnected parameters in neural networks, it is impractical to distill the system's overall behavior into a clear set of rules. This inherent difficulty in interpreting how LLMs process language and arrive at conclusions persists even for their creators, who cannot, for example, explain why LLMs respond to emotional pleas.[4] Reducing scale could improve interpretability but most likely at the cost of capability, as scale enables unique emergent behaviors. And so there are intrinsic trade-offs—between scale, capability, and interpretability—meaning opacity is a structural feature, not a bug, of LLMs.

Exhibit 2
LLMs are Progressing at Twice the Speed of Moore's Law



As of 30 April 2024
Source: Anson Ho et al.

Small Language Models

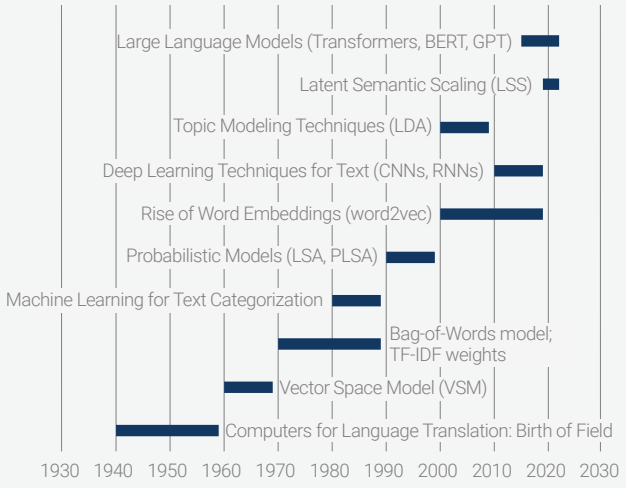
"I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worthwhile."

– Warren Weaver, "Translation," 1949.[5]

Three-quarters of a century ago, Warren Weaver's seminal 1949 memorandum laid the groundwork for machine translation, statistical language models, and modern natural language processing (NLP) techniques. Comparing translation to code-breaking, Weaver posited that the "semantic difficulties" of translation, such as multiple meanings, could be solved computationally.

The 1960s saw the birth of dictionary-based quantitative text analysis with the creation of *The General Inquirer* by Stone et al. in 1966. This program searched documents for words registered in a pre-defined keyword list or "dictionary." From these humble beginnings, the following decades witnessed the development of an extensive suite of text processing techniques, which we have collectively named Small Language Models (Exhibit 3).

Exhibit 3
SLMs Have Been Refined over Decades



As of 30 April 2024
Source: Lazard

Unlike their large cousins, SLMs do not hallucinate, consistently producing the same outcomes given the same input. They also operate as "glass boxes," with a much smaller number of parameters. One such SLM, developed by Kohei Watanabe, a member of the Lazard Quantitative Equity team, is Latent Semantic Scaling (LSS). LSS is a semi-supervised document scaling technique that locates documents in any language on user-defined dimensions.[6]

To achieve this, LSS combines two elements: the unsupervised Latent Semantic Analysis (LSA) technique and user-provided "seed words." LSA estimates the semantic proximity between words in the data, while the seed words define the dimensions of interest. By measuring the semantic proximity of words to the seed words instead of relying solely on word frequencies, LSS provides a highly interpretable, theoretically grounded, and precise answer to a given question in any language. In the next section, we demonstrate its use in practice via a case study on managing risk emanating from the Chinese housing market.

Case Study: China's Housing Market

In August 2020, the Chinese Government announced its "Three Red Lines" policy, imposing stricter financial requirements on real estate developers to deleverage the sector, tame bubbly property prices, and mitigate systemic financial risks.

China's housing sector has been mired in a prolonged slowdown ever since, with numerous property developers, including Evergrande, Country Garden, and Shimao Group, defaulting on their debts or facing liquidation. This distress threatens household wealth and has been a key driver of the country's economic slowdown, weighing on the broader emerging markets index for the past three years (Exhibit 4).

Exhibit 4
Emerging Equity Markets Have Been Held Back by China

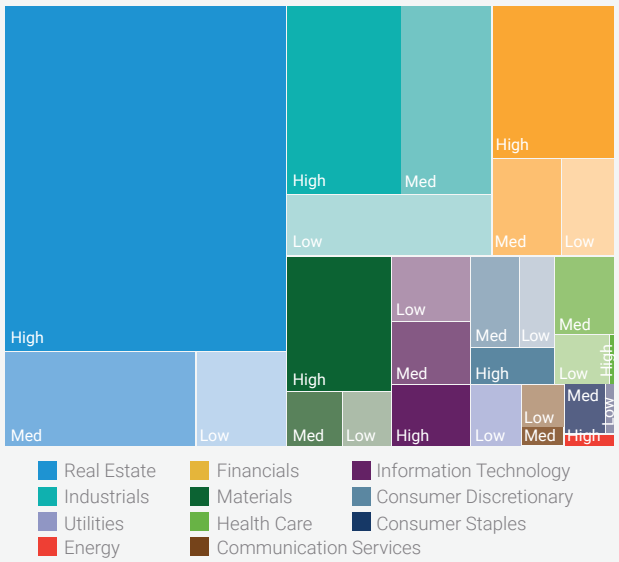
As of 10 April 2024
Source: Bloomberg, MSCI

To help us navigate this "slow-motion financial crisis,"[7] we applied LSS to annual reports filed in Mandarin (Exhibit 5). By analyzing text in its original form, we can gain a deeper understanding of the nuances, cultural references, and contextual meanings that are often lost in translation. The analysis covered 25,000 documents comprising of 6.5 million pages. Notably, this was achieved using only a handful of parameters and without requiring fluency in Mandarin, because semantic context is extracted statistically.

Exhibit 5
CITIC, an Industrial Conglomerate, Comments on the Chinese Housing Market

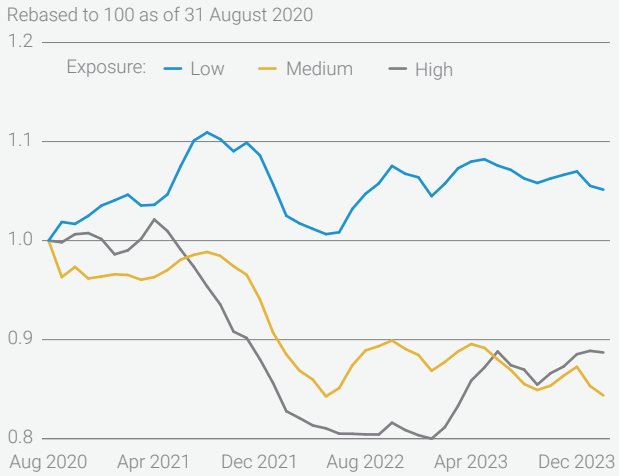
As of 30 April 2024
For illustrative purposes only.
Source: Lazard, company filings

Exhibit 6
Chinese Housing Market Exposure beyond the Real Estate and Finance Sectors



As of 30 April 2024
For illustrative purposes only.
Source: Lazard, company filings

Exhibit 7
LSS Accurately Captures Exposure to the Housing Market Theme

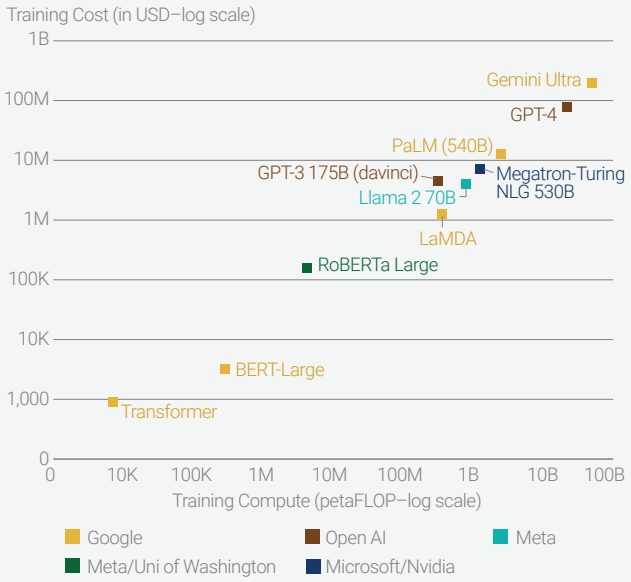


As of 30 April 2024
Source: Lazard, company filings

Exhibit 6 shows the results. Unsurprisingly, the topic is front and center for property developers and the banks that provide them with loans. Industrial companies also feature prominently, with almost twice as many as financial firms having medium-to-high exposure. These include companies involved in engineering, procurement, and construction (EPC) projects, machinery manufacturers such as excavator makers, and firms providing housing-related services like waste management. Facing slowing domestic sales, many are increasingly turning to export markets, leading to anti-dumping probes from both the UK and European Commission.[8,9]

Crucially, the stock returns of companies outside the real estate and financial sectors have aligned with their degree of exposure to the housing market, as estimated using LSS. Those less exposed have seen better stock performance compared to their more exposed counterparts and the benchmark (Exhibit 7). In addition to being transparent—we can completely unpack and fully understand LSS outputs—this level of precision comes at a fraction of the cost if we were to attempt the same exercise using an LLM (Exhibit 8).

Exhibit 8
Estimated Training Cost and Compute of Select AI Models



As of 30 April 2024
Source: Lazard, company filings

Conclusion

In a world captivated by large language models, it is easy to overlook the value of their smaller cousins. Small language models drawing on techniques refined over 75 years, offer reliability and interpretability—qualities often lacking in LLMs. While LLMs dazzle with their scale and scope, their opacity and tendency to hallucinate limit their utility in the realm of investing. SLMs, in contrast, deliver consistent, tractable results that are especially suited to highly specific questions, as our case study on China's housing market has shown. And so, as the AI race accelerates, investors and researchers should remember that big is not always beautiful.

This content represents the views of the author(s), and its conclusions may vary from those held elsewhere within Lazard Asset Management. Lazard is committed to giving our investment professionals the autonomy to develop their own investment views, which are informed by a robust exchange of ideas throughout the firm.

Notes

1 Anson Ho et al. (2024). Algorithmic progress in language models. Retrieved from https://doi.org/10.48550/arXiv.2403.05812

2 The Guardian (2024). Air Canada ordered to pay customer who was misled by airline's chatbot. Retrieved from https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit

3 Reuters (2023). New York lawyers sanctioned for using fake ChatGPT cases in legal brief. Retrieved from https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/

4 Cheng Li et al. (2023). Large Language Models Understand and Can be Enhanced by Emotional Stimuli. Retrieved from https://doi.org/10.48550/arXiv.2307.11760

5 Warren Weaver (1955). Translation in Locke, W.N. and Booth, A.D. (Ed.), "Machine translation of languages: fourteen essays" (pp. 15-23). Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass.

6 Kohei Watanabe (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. "Communication Methods and Measures," 15(2), 81–102.

7 Financial Times (2022). China's property crash: 'a slow-motion financial crisis'. Retrieved from https://on.ft.com/3UiFAEF

8 Trade Remedies Authority (2023). TRA opens new investigations into excavators from China. Retrieved from https://www.gov.uk/government/news/tra-opens-new-investigations-into-excavators-from-china

9 Council of the European Union (2023). Notice of initiation of an anti-dumping proceeding concerning imports of mobile access equipment (MAE), originating in the People's Republic of China. Retrieved from https://eur-lex.europa.eu/eli/C/2023/783/oj

Important Information

Published on 28 May 2024.