Title: Unleashing the Economic Potential of Large Language Models: The Case of Chinese Language Efficiency

Future Large language model efficiencies may lead to regional economic advantages due to fundamental linguistic disparities.

Barnas Monteith

THInc AI Group

Corresponding Author:

Michael Sung

IDFI Laboratory

Zhejiang University

michaelsung@intl.zju.edu.cn

Abstract

Abstract: Large language models (LLMs) have revolutionized the way we interact with technology, and ChatGPT, based on the GPT-3.5 architecture, has garnered significant attention for its exceptional performance. The widespread adoption of ChatGPT by over 100 million users within three months, generating 1.8 billion website visitors per month, highlights its versatility and appeal. The economic impact of LLMs extends beyond businesses, empowering developers and entrepreneurs to create innovative products and services. This paper explores the potential of LLMs, specifically focusing on the efficiency and advantages offered by the Chinese language.

The Chinese language possesses unique features, including compactness, polysemy, contextual nature, homophony, classical roots, and an ideographic nature, which contribute to its expressiveness. Recent advancements in contextual information processing and ideograph encoders have further improved the performance of Chinese language-based generative AI. Additionally, China's linguistic advantage provides economic opportunities in various areas. China's massive market size and user base make it an attractive target for language-based applications built on LLMs. The availability of abundant training data enhances the accuracy and contextual understanding of Chinese language models, giving Chinese researchers and developers an edge in LLM development. Furthermore, the efficiency of the Chinese language enables language-specific applications, catering to Chinese users in sectors such as e-commerce, customer service, finance, education, and healthcare. The cultural context and localization of Chinese language models foster better user engagement and satisfaction.

The economic advantages of the Chinese language in the LLM landscape position China for future opportunities. With the rapid growth of internet users in China and the abundance of Chinese language content available for training, Chinese LLMs have the potential to surpass English LLMs in scalability and performance. Leveraging the efficiency and unique characteristics of the Chinese language, China can drive adoption, create culturally sensitive experiences, and gain a competitive edge in developing advanced language-based applications.

As large language models continue to evolve, their impact on technology and various industries is expected to be transformative. The Chinese language's efficiency and regional economic advantages highlight the potential for continued growth and innovation in the context of LLMs, shaping the future of human-computer interaction and opening new avenues for economic development.

Introduction:

In recent years, large language models have emerged as a transformative technology, revolutionizing the way we interact with computers and the internet. These models, powered by advanced deep learning algorithms, have demonstrated unprecedented capabilities in understanding and generating human language. Among the notable breakthroughs in this domain, ChatGPT, based on the GPT-3.5 architecture, has gained significant attention due to its exceptional performance and versatility.

One measure of the success of large language models / generative AI is the volume of users they attract. With the exponential growth of internet users and the increasing demand for sophisticated natural language processing systems, the adoption of ChatGPT has soared to new heights. Its user base comprises individuals, businesses, and organizations seeking solutions for various purposes, such as customer support, content generation, virtual assistants, and even creative writing. The widespread appeal and adoption of ChatGPT can be attributed to its ability to comprehend context, generate coherent responses, and adapt to specific user needs. These features have thus attracted a community of over 100 million users, within a three month period, at the beginning of 2023, currently generating 1.8 billion website visitors per month. This rate of growth is substantial, as it exceeds that of highly popular social media platforms within the US, such as TIktok and Instagram. OpenAI, the maker of ChatGPT, has estimated revenues of 200 million $ in 2023, projecting over $1 billion by 2024. It is also important to note that this figure does not include related or derivative applications that now depend on ChatGPT for core functionality.

The widespread economic success of large language models like ChatGPT in markets beyond chatbots and related online generative NLP AI applications has been remarkable. Their deployment has paved the way for numerous applications, enabling businesses to enhance their operations, improve customer experiences, and increase productivity. ChatGPT, with its language generation capabilities, has empowered companies to automate tasks that were previously time-consuming and resource-intensive, such as answering customer queries, generating personalized recommendations, and providing tailored content. By leveraging ChatGPT, organizations have not only achieved cost savings but have also witnessed improved customer satisfaction, increased revenue, and gained a competitive edge in the market.

Moreover, the economic impact of large language models extends beyond businesses. Developers and entrepreneurs have seized the opportunities provided by these models to create innovative products and services. The availability of APIs and developer tools has democratized access to this technology, allowing individuals and startups to build applications without requiring extensive expertise in machine learning. This accessibility has led to the emergence of a vibrant ecosystem, with developers building chatbots, language-based games, educational tools, and more. This flourishing ecosystem has not only fostered technological advancement but has also contributed to economic growth and job creation.

The recent success of large language models, exemplified by the popularity and economic impact of ChatGPT, can be attributed to their ability to understand and generate human language at an

unprecedented level. The widespread adoption of ChatGPT by diverse user groups, ranging from individuals to businesses, showcases its broad appeal and versatility. Furthermore, the economic success associated with large language models is evident through the numerous business applications and the entrepreneurial opportunities they have unlocked. As we move forward, the continued advancement and refinement of large language models are expected to shape the way we interact with technology and revolutionize various industries, and lead to new, innovative derivative business models, entirely driven by the LLM backend.

Generative AI/LLMs are fundamentally enabled by an enhanced ability to determine semantic equivalence, which is in part due to improved transformer AI methodologies, but also arguably primarily through scaling. It has been thought that the increased scale of generative LLM models is the causal correlation for their improved accuracies/performance over time and thus, overall usefulness to end users. However, new research indicates that scaling of input training data, may have a far greater effect of the efficacy of the model, than the number of parameters (with GPT3 at around 175 billion parameters, and GPT4 rumored to have over 1 trillion parameters), due to model error, computational power requirements and overall data collection and training costs.

A recent report published by Dynomight addresses the issue of scaling in LLM's. (Dynomight 2023). While in recent years, many researchers have been focused on increasing the number of parameters in LLMS (with GPT-3, Gopher and PaLM, 80% of error was due to limited training data and not model size). By using the same amount of computational power and cost on training LLM models with additional data, but not necessarily focusing on increasing parameters, the final models will be far more efficient, with a much smaller cost and carbon footprint. With substantially improved ROI, this will lead to greater proliferation of such models.

Chinese language linguistic advantages

Chinese, as a language, has several fundamental semantic efficiency advantages that contribute to its unique features and expressive power. Some of these advantages include:

Compactness: Chinese characters, known as Hanzi, are logographic, meaning each character represents a word or a morpheme. This allows for a high degree of information density, as a single character can convey meaning that might require multiple alphabetic letters or syllables in other languages. This compactness enables Chinese speakers to express complex ideas concisely.

Polysemy: Chinese characters often have multiple meanings depending on the context. While this can occasionally lead to ambiguity, it also allows for nuanced and layered expressions. The flexibility of interpreting a character in various ways adds depth to the language and facilitates the expression of abstract concepts.

Chang, et al (2021) have developed a new method for addressing the ambiguities of polysemy, utilizing BERT (bi-directional encoder representations from transformers) to improve contextual information, allowing for greater performance of the word embedding approach. This, and similar approaches, will further improve the ability for Chinese language-based generative AI / LLLMs to attain high level performance and accuracy, in the near future.

Contextual nature: Chinese relies heavily on context to convey meaning. Due to the lack of inflections and verb conjugations, the same words can be used in different grammatical roles depending on the

sentence structure and context. This contextual nature enhances efficiency by reducing the need for explicit grammatical markers and enabling speakers to convey information economically.

Homophony: Chinese has a relatively limited number of distinct syllables compared to other languages. This results in a higher degree of homophony, where different characters share the same pronunciation. While this can present challenges for understanding spoken language without accompanying characters, it also enables various wordplay techniques, such as puns, rhymes, and allusions, enhancing the poetic and artistic aspects of the language. While homophony does not necessarily need to be addressed for LLMs, related NLP applications such as TTS (text to speech), will need to address homophony. It is thought that a similar approach to the polysemy solution mentioned above, could be used to improve contextual inference of word meanings.

Classical roots: Chinese characters have a long historical tradition, and many of them are derived from ancient pictographs and ideographs. This historical depth allows Chinese to capture concepts and abstract meanings with visual cues. Additionally, Chinese characters can retain their meaning over time, even when pronounced differently, preserving semantic information and providing a sense of continuity with the past.

Ideographic nature: Chinese characters often contain visual or symbolic elements that provide clues about their meanings. For example, the character for "tree" (木) includes a pictorial representation of a tree, forest, or senlin in Chinese is represented by the character: 森林 – which contains multiple representations of the character for the English word “tree.” This logical, idiographic nature can aid in understanding and memorization, as the visual elements can serve as mnemonic devices, reinforcing the semantic efficiency of the language.

A study involving the development of a recurrent neural network using radical-level ideograph encoders for sentiment analysis based on Chinese and Japanese languages has demonstrated results comparable to modern word vectorization / embedding approaches, while at the same time, being cost effective to train. (Ke and Hagiwara, 2017)

It's important to note that while the Chinese language possesses these semantic efficiency advantages, it also poses challenges in terms of character memorization, complex writing system, and the need for extensive vocabulary knowledge. Nonetheless, these features contribute to the rich and expressive nature of the Chinese language, and LLM’s / generative AI can help to expand greater global adoption of Chinese, and enhance the business use of the Chinese language, through the easing of these human-centric challenges, as an assistive technology, with a track record of continuous improvement over time.

Chinese language efficiency may lead to regional economic advantage in a future increasingly reliant on generative AI & LLMs.

The efficiency of the Chinese language, in terms of both its structure and the sheer number of speakers, presents a significant opportunity for China in the context of large language models. As large language models continue to evolve and gain prominence, China's linguistic advantage positions it for potential future economic advantages in several key areas.

Market Size and User Base: Chinese is the most widely spoken language in the world, with over 1.4 billion native speakers. This vast user base provides a substantial market for applications built on large language models. As the adoption of language-based technologies grows, China's market size ensures a

ready audience and a potentially lucrative customer base. This advantage allows Chinese developers and businesses to target domestic consumers effectively and scale their products or services to a massive population.

Data Availability: Large language models require significant amounts of high-quality training data to achieve optimal performance. With a vast user base and a digital landscape that generates substantial linguistic data, China possesses a wealth of Chinese language resources. This availability of diverse and abundant training data gives Chinese researchers and developers an edge in training models specifically tailored for the Chinese language. This advantage can lead to improved accuracy and better contextual understanding in Chinese language models, further strengthening China's position in the development and deployment of large language models.

A relatively recent comparative language vectorization study by Graves et al (2018) found that among 157 languages, there were substantial differences in the ability to train a system to learn new word vectorizations, using large scale noisy data from the Internet (in this study, Common Crawl and Wikipedia were the primary data sets used). It was found that the influence of training data quality and type was significant, and in the case of Finnish and Hindi (which were trained on data sets containing smaller Wikipedia training sets), performance improvements were observed (+23.5 improvement), and Chinese also demonstrated comparably improved performance (+17.8). Languages with greater availability of web data (from which higher quality subsets of data can be wrangled/curated), may therefore show greater performance when used in the training of future non-English LLMs.

Zhang et al (2021) have demonstrated a successful attempt at developing a large scale Chinese Pre-trained language model (CPM), with 2.6 billion parameters and 10GB of training data, showing strong performance at multiple NLP tasks including few shot and zero-shot learning.

According to a report by Visual Capitalist, both Chinese and English are among the top 20 languages used on the Internet (from which all LLM models scrape their training data), with English comparably representing 16.2% of the speaking population and Chinese representing 14.3%. However, today, Chinese is only used by 1.4% of the world's top 10 million websites. Yet, China has among the fastest growth rates in the world, in terms of Internet usage and content production. According to Statista, the number of Internet users in China grew by 35 million, year over year, as of December 2022, growing from 564 million people in 2012 to 1.06 B people by 2022, doubling in just a decade, and with a penetration rate of over 75% of the entire Chinese population having Internet access. It is expected that the growth of penetration and total Internet users, and thus, available training content, will continue to grow rapidly in the near future. According to Textmaster, the number of English users online is 1186 million, having grown 743% from 2000 to 2020, but not nearly at the same rate as Chinese language users, which has risen 2,600% in that same time. This means that the amount of available Chinese language content for scraping and training may soon surpass that of English in the near future. (Textmaster).The scaling potential for Chinese LLM's will far exceed that of English in the long term (Statista) (Bhutada).

Language-Specific Applications: The efficiency of the Chinese language, characterized by the use of characters instead of phonetic alphabets, offers unique possibilities for applications that rely heavily on language processing. Tasks like machine translation, sentiment analysis, content generation, and voice recognition can leverage the inherent structure and richness of the Chinese language. With large

language models, China can develop advanced language-based applications that cater specifically to Chinese users, providing them with highly accurate and contextually relevant solutions. These applications can span across various sectors, including e-commerce, customer service, finance, education, healthcare, and more, potentially giving Chinese businesses a competitive edge in these domains.

Cultural Context and Localization: Language is intricately connected to culture, and the Chinese language is deeply rooted in Chinese culture and society. Large language models trained specifically for Chinese can better understand the cultural nuances, idioms, and references that are unique to Chinese society. This contextual understanding enables the development of localized and culturally appropriate applications, facilitating better user engagement and satisfaction. By providing highly relevant and culturally sensitive experiences, China can foster a stronger bond between its people and technology, driving adoption and economic growth.

In a 2018 study by Li, et al., Chinese analogical reasoning was studied. Using both semantic and morphological relations of Chinese words. Analogical reasoning is now possible for generative AI/LLM due to modern methods of analyzing and organizing "linguistic regularities" using word embedding techniques common in NLP applications. In this study, in order to investigate semantic knowledge reasoning, semantic relationships of words were classified by clustering relationships into categories: geography, nature, history and people. These categories are highly linked to culture, and it was observed that there are specific nuances in the relationships between Chinese language and culture that differ from other languages such as English, based on the research team's investigation of Chinese lexical knowledge. Cultural context will be incorporated into considerations for future NLP technologies, including LLMs. It is thought that this approach will improve the final model's ability to parse semantic relationships/equivalence in future generations of these models, leading to greater accuracy and performance of the models.

In summary, the efficiency and vast user base of the Chinese language provide China with potential economic advantages in the context of large language models. The size of the market, coupled with the availability of diverse training data and the unique characteristics of the language, positions China to leverage large language models for developing applications tailored to Chinese users. By focusing on language-specific applications and considering cultural context, China can gain a competitive edge in various sectors and drive economic growth through innovative language-based technologies.

Recently, a number of Chinese companies have begun their own LLM development projects, including Baidu, Flytek, Tencent and Alibaba (TongyiQianwen). (The Hindu Bureau, April, 2023). Further, there have been a number of announcements with regard to computing infrastructure improvements, to facilitate the training of these LLM models (for instance, Tencent's investment in a new supercomputing cluster specifically for this purpose (PingWest 2023). It is expected that China's investments into quantum computing and improvements in classical computing semitech will further accelerate the speed and development of new models for generative AI. For instance, China's Jiuzhang 2 has been reported to achieve processing speeds up to 1 million times faster than that of Google's Sycamore Quantum computing architecture.

References

Bergan, B. (2021, October 27). China's New Quantum Computer Has 1 Million Times the Power of Google's. *interestingengineering.com*. https://interestingengineering.com/innovation/chinas-new-quantum-computer-has-1-million-times-the-power-of-googles

Bhutada, G. (2021, July 29). *Visualizing the Most Used Languages on the Internet*. Visual Capitalist. https://www.visualcapitalist.com/the-most-used-languages-on-the-internet

Chang, Y., Kong, L., Kejia, J., & Meng, Q. (2021). Chinese named entity recognition method based on BERT. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*. https://doi.org/10.1109/icdsca53499.2021.9650256

Dynomight. (2023, March 8). First-principles on AI scaling. https://dynomight.net/scaling/#:~:text=Scaling%20laws%20say%20that%20with,any%20new%20breakthroughs%2C%20just%20scale

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., &Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Language Resources and Evaluation*. Springer Science+Business Media. https://infoscience.epfl.ch/record/253313/files/1802.06893.pdf

The Hindu Bureau. (2023, April 11). *China's Alibaba reveals its large language model TongyiQianwen*. https://www.thehindu.com/sci-tech/technology/china-alibaba-reveals-large-language-model-llm-tongyi-qianwen/article66724187.ece/amp

Ke, Y., & Hagiwara, M. (2017). Radical-level Ideograph Encoder for RNN-based Sentiment Analysis of Chinese and Japanese. In *Asian Conference on Machine Learning* (Vol. 77, pp. 561–573). http://proceedings.mlr.press/v77/ke17a/ke17a.pdf

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). *Analogical Reasoning on Chinese Morphological and Semantic Relations*. https://doi.org/10.18653/v1/p18-2023

PingWest. (2023, April 14). Tencent launches supercomputing cluster to aid LLM training in China. https://en.pingwest.com/w/11603

Statista. (2023, April 19). *Number of internet users in China 2022.*https://www.statista.com/statistics/265140/number-of-internet-users-in-china

Zameo, S. (2021, February 11). *Which language will dethrone English on the internet? | TextMaster*. The International Expansion Blog. https://www.textmaster.com/blog/language-dethrone-english-internet

Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., Qin, Y., Su, Y., Ji, H., Guan, J., Qi, F., Wang, X., Zheng, Y., Zeng, G., Cao, H., Chen, S., Li, D., Sun, Z., Liu, Z., . . . Sun, M. (2021). CPM: A large-scale generative Chinese Pre-trained language model. *AI Open*, *2*, 93–99. https://doi.org/10.1016/j.aiopen.2021.07.001

Zhou, J., Wang, J., & Liu, G. (2019). *Multiple Character Embeddings for Chinese Word Segmentation*. https://doi.org/10.18653/v1/p19-2029