Poster Abstract: On the Accuracy and Robustness of Large Language Models in Chinese Industrial Scenarios

Zongjie Li*, Wenying Qiu†, Pingchuan Ma*, Yichen Li*, You Li†, Sijia He†, Baozheng Jiang†
Shuai Wang*¶, Weixi Gu†¶
*Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China
†China Academy of Industrial Internet, Beijing, China
zligo, pmaab, ylipf, shuaiw@cse.ust.hk
qiuwenying, liyou, hesijia, jiangbaozheng, guweixi@china-aii.com

***Index Terms*—Large language model, AI reliability**

I. Study Purpose

Recent studies have demonstrated that large language models (LLMs) exhibit exceptional performance across various natural language processing tasks, rivaling or even exceeding human competencies in certain areas [1]–[5]. Typically, LLMs undergo pre-training on extensive text corpora, usually using billions of tokens to develop a foundational model. To better align LLMs with human preferences and directives or to fulfill specific application needs, methods such as supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and direct preference optimization (DPO) have been introduced and demonstrated to be effective. These advancements facilitate more intuitive and efficient human-AI interactions. However, the substantial resource requirements throughout the training process pose challenges for individual users and smaller organizations.

Despite the success of LLMs, most are trained with English corpora, and the majority of LLMs are designed for English users. For example, prominent open-source LLMs such as LLaMA2 are predominantly trained on English corpora with limited ability in understanding non-English texts, while advanced multilingual models like GPT-4 are often proprietary and unavailable publicly. Therefore, to cater to the Chinese market and adhere to legal regulations, many Chinese LLM vendors adopt localization strategies, providing "local" LLM specially designed for Chinese by pretraining models on large-scale Chinese datasets or fine-tuning them using supervised instruction corpora.

In recent times, there has been a growing trend of integrating LLMs into manufacturing production pipelines [6]. However, the high threshold for training resources and corpora, along with the lack of testing for non-English LLMs, raises concerns regarding their accuracy and robustness [7]. In the context of industrial manufacturing, accuracy is crucial to prevent potential catastrophic defects that could result in significant losses [8]. Additionally, robustness is vital as manufacturing models frequently operate under constrained conditions, necessitating deterministic outputs rather than conversational ones. Despite the successes in dialogue applications, industrial manufacturing organizations remain hesitant to adopt LLMs within critical production systems due to these challenges. Ensuring sufficient accuracy and reliability to gain stakeholder trust and mitigate risks remains a significant hurdle for the widespread implementation of LLMs in the manufacturing sector.

¶ Corresponding authors.

II. Study Methods

To address this gap, we present a comprehensive empirical study that assesses the accuracy and robustness of local LLMs in industrial scenarios. The study was conducted through collaboration between nine top industrial research teams in China. We carefully considered relevant administrative regulations and laws and manually curated 1,200 industry-specific problems across eight industrial sectors to assess accuracy. To further quantify the robustness of the LLMs, we developed a metamorphic testing framework with industry-oriented relations. This framework evaluated four stability categories with eight abilities through 12,431 variants. In total, we evaluated eight different local LLMs developed by Chinese vendors and two different LLMs developed by global vendors. Overall, our study aims to address the following four research questions (RQs): **RQ1:** To what extent are LLMs accurate in Chinese industrial scenarios? **RQ2:** How robust are LLMs across different Chinese industrial scenarios?

We present the selected industrial sectors and the problems in Table I. The industrial base of a mature country typically comprises thousands of sectors across various industries such as manufacturing, mining, energy, and materials [9]. China, for instance, boasts one of the most diverse and extensive industrial bases globally, with tens of thousands of sectors [10]. Among these numerous sectors, eight key industries can be identified that have high production values, utilize advanced automation, and have the potential for further integration of LLMs into their industrial chains. We list these sectors below:

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00042

TABLE I
FOUR CATEGORIES AND EIGHT ABILITIES OF THE STABILITY CATEGORY.

Name	Stability	Description	Original texts	Equivalent variants
Magnitude change	Numeric	Equivalent substitution of data outlines	100cm	1m
Digital precision	Numeric	Changing data precision	3.7m	3.70m
Synonyms	Grammar	Replacement with industry-specific synonyms	USB	U-PAN
Order	Grammar	Swapping the order of options	A. five B. six	A. six B. five
Logic	Grammar	Reversal the logic of the question	You should touch xx	You should not touch xx
General context	Context	Add background for testing Industrial sectors	[Q]	In xxx industry, [Q]
Security context	Context	Add security instruction	[Q]	Consider xxx law, [Q]
Irrelevant content	Others	Add irrelevant option	A—B—C—D	A—B—C—D—E

Electronic equipment manufacturing: Fabrication of electronic devices, components, specialized materials, and other electronic components. Equipment manufacturing: Production of metal products, general equipment, specialized equipment, and automobiles. Iron and steel industry: Ironmaking, steelmaking, steel rolling and processing, ferroalloy smelting. Mining industry: Extraction of coal, oil, natural gas, ferrous metals, nonferrous metals, and other minerals. Power industry: Electricity generation, transmission, heat production, and distribution. Petrochemical industry: Petroleum refining and processing, manufacturing of chemical feedstocks and products, plastic goods, rubber goods.

The industrial sectors examined represent fundamental components in the manufacturing chain, as well as the industries with the most extensive real-world applications. These sectors are selected due to their importance in industrial processes and ubiquity across supply chains. **Notably, the research focuses exclusively on civilian industries and does not encompass any sectors related to national security or defense.**

III. STUDY FINDINGS

We obtain many findings, and some of the major ones include: The accuracy of all evaluated LLMs remains insufficient (under 60%) for deployment in industrial applications, and it varies substantially across industrial areas, especially for less mature models. Most LLMs tend to be more robust on the variants whose corresponding seed question is correctly answered than those answered wrongly. The robustness scores vary across industrial sectors, and local LLMs are generally worse than global LLMs, which may be attributed to the quantity and quality of available data on the Internet. LLM robustness differs significantly across abilities. Global LLMs demonstrate greater robustness to logical perturbations, while top local LLMs better understand Chinese industrial terminology.

Our findings provide guidance for developing LLMs that better serve non-English (Chinese) users in industrial applications. They will assist platform engineers and enterprises in improving local LLMs for manufacturing. To summarize, this paper makes the following contributions: leftmargin=*

- We perform the first comprehensive study on the accuracy and robustness of LLMs in Chinese industrial scenarios.
- We collect the first benchmark of industry-specific problems in Chinese.
- We propose a metamorphic testing framework with industrial metamorphic relations to assess robustness in Chinese industrial scenarios.
- We systematically evaluate 10 LLMs from 9 different vendors, and compare the local LLMs with global LLMs in terms of multiple abilities.
- We point out the implications of our findings and suggest possible improvements for the development and usage of LLMs in Chinese industrial scenarios.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[3] Z. Li, C. Wang, S. Wang, and G. Cuiyun, "Protecting intellectual property of large language model-based code generation apis via watermarks," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*. ACM, 2023.

[4] C. Wang, Z. Li, Y. Peng, S. Gao, S. Chen, S. Wang, C. Gao, and M. R. Lyu, "Reef: A framework for collecting real-world vulnerabilities and fixes," *arXiv preprint arXiv:2309.08115*, 2023.

[5] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang, "Insightpilot: An llm-empowered automated data exploration system," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 346–352.

[6] "google-palm2-security," https://cloud.google.com/blog/products/identity-security/security-ai-next23.

[7] Z. Li, C. Wang, Z. Liu, H. Wang, D. Chen, S. Wang, and C. Gao, "Cctest: Testing and repairing code completion systems," in *Proceedings of the 45th International Conference on Software Engineering*, ser. ICSE '23, 2023, p. 1238–1250.

[8] "Error in Bard Demo," https://edition.cnn.com/2023/02/08/tech/google-ai-bard-demo-error/index.html.

[9] L. Bernstein, J. Roy, K. C. Delhotal, J. Harnisch, R. Matsuhashi, L. Price, K. Tanaka, E. Worrell, F. Yamba, Z. Fengqi *et al.*, "Industry," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2007.

[10] B. of Operation Monitoring, M. o. I. Coordination, and I. Technology, "China industrial economic operation report 2022," http://lwzb.stats.gov.cn/pub/lwzb/bztt/202306/W020230605407820366191.pdf, 2023.