Bio-Design and Manufacturing (2024) 7:819–822
https://doi.org/10.1007/s42242-024-00307-0

NEWS & VIEWS

Evaluation of large language models for the classification of medical device software

Yu Han[1] · Aaron Ceross[1] · Florence Bourgeois[2,3] · Paulo Savaget[1] · Jeroen H. M. Bergmann[1,2,4]

Received: 17 June 2024 / Accepted: 20 June 2024

Amidst the rapidly expanding integration of large language models (LLMs) across various sectors (ranging from everyday applications to specialized fields demanding stringent regulatory adherence), our investigation seeks to determine how well these models can support medical device software classification. Medical device classification functions to systematically categorize devices according to their designated use, associated risk levels, and requisite regulatory oversight, thereby providing a structured framework for ensuring safety and efficacy as mandated by regulatory authorities. This classification paradigm is integral to the regulatory landscape, guiding the oversight mechanisms that ensure medical devices to meet predefined safety standards before they can be utilized in healthcare settings [1]. Most national frameworks utilize a risk-based classification scheme [2]. For instance, in the USA, a Class III device, which necessitates the most rigorous oversight, must undergo thorough review via the pre-market approval (PMA) process [3]. Nonetheless, the classification criteria exhibit considerable variation owing to the disparate regulatory environments across countries, presenting challenges to global compliance and harmonization efforts [4]. Our analysis spans an extensive range of devices from important international markets including China, the USA, and Europe. It provides an assessment of the capacities of LLMs to support medical device software classification.

Trained on extensive datasets sourced from the Internet, literature, and an array of textual materials, LLMs represent a pivotal advancement in artificial intelligence (AI). These models excel in grasping complex language patterns, context, and semantics, operating on predictive algorithms that determine the likelihood of subsequent words based on existing sequences. This capability provides LLMs with an unprecedented capacity for comprehending, producing, and navigating human language, positioning them as a potential suitable tool to support regulatory decision making.

Regulatory affairs, captured by the International Medical Device Regulators Forum (IMDRF) [5], ensure that medical products conform to the varied and rigorous regulatory standards set by jurisdictions worldwide. Companies are tasked with navigating this multifaceted terrain, balancing the submission of comprehensive dossiers against the demands of scientific evaluations by regulatory bodies. Challenges arise when determining the risk classification of devices in development in order to tailor evidence generation. In the USA, devices are systematically classified into three main categories (Class I, Class II, and Class III), each specifying the degree of regulatory scrutiny required, which in turn influences the approval process and strategies for market entry. In contrast, the European Union (EU) employs a more detailed classification framework, segmenting devices into four categories (Class I, Class IIa, Class IIb, and Class III). Misclassification can lead to major operational and financial setbacks, such as unnecessary clinical trials or market access denials [6]. Recognizing the appropriate classification early in the development process is essential for manufacturers to tailor evidence generation and compile robust regulatory submissions. Each classification level, indicative of the device's potential risk and requisite oversight, prescribes specific reg-

✉ Jeroen H. M. Bergmann
jeroen.bergmann@eng.ox.ac.uk

Yu Han
yu.han@eng.ox.ac.uk

Aaron Ceross
aaron.ceross@eng.ox.ac.uk

Florence Bourgeois
florence.bourgeois@childrens.harvard.edu

Paulo Savaget
Paulo.Savaget@eng.ox.ac.uk

1 Department of Engineering Science, University of Oxford, Old Road Campus, Oxford OX3 7QD, UK

2 Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA

3 Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA, USA

4 Department of Technology and Innovation, The University of Southern Denmark, Odense, Denmark

ulatory pathways. These pathways differ markedly across various international jurisdictions, impacting not only the speed to market, but also the marketability and competitive positioning of the device in different markets.

Recent studies underscore the increasing complexity of regulatory affairs and the multifaceted challenges it introduces [7, 8]. This complexity is further exacerbated by the rapid pace of technological innovation and the expanding variety of medical devices. In this context, the development of an accurate and adaptable classification methodology emerges as a crucial endeavor, one that could greatly benefit from the integration of AI. Traditional machine learning techniques have previously been applied to tackle classification challenges within this domain [9, 10], with a recent study achieving a weighted accuracy of 82% using support vector machine (SVM) [11]. Comparably, the ability to use LLMs globally for medical device classification is promising; however, no data tested exist on the accuracy.

We queried several LLMs to assess the feasibility for device classifications, comparing these with the official classifications from respective regulatory databases. We selected seven global cutting-edge LLMs for our study: (1) GPT-3.5, (2) Gemini, (3) Qwen, (4) ERNIE Bot, (5) Baichuan, (6) ChatGLM, and (7) Mistral Large. These models, hailing from the USA, the EU, and China, represent the state of the art of generative AI. These LLMs are based on the Transformer model [12], indicating the importance of model size in their ability to grasp and produce complex text. Among them, ERNIE Bot, with 260 billion parameters, stands out as one of the largest, suggesting a superior capability in handling nuanced textual interpretations. However, a larger parameter count also means a greater demand for computational resources.

The dataset consisted of 130 distinct medical software products: 66 devices from the National Medical Products Administration (NMPA) of China, 29 devices from the U.S. Food and Drug Administration (FDA) [13], as well as 35 devices from EU database guided by the listings in Ref. [14]. This dataset classification result has been served as the golden standard against which we could evaluate the performance of the LLMs. The whole dataset is shown in Supplementary Information.

For each product evaluated, the generic name of the medical devices was used rather than their brand names. Generic names, also referred to as non-proprietary names, encompass descriptions that articulate the device's intended use, underlying technology, and pertinent details such as the device's indications, functionalities, and related disease information [15]. This approach aligns with the idea that a classification should be based on the device description and not on any brand name. The classification of each device by the authorities is predicated primarily on the text delineating its intended purpose, as this serves as the foundational criterion for regulatory assessment and subsequent categorization [11].

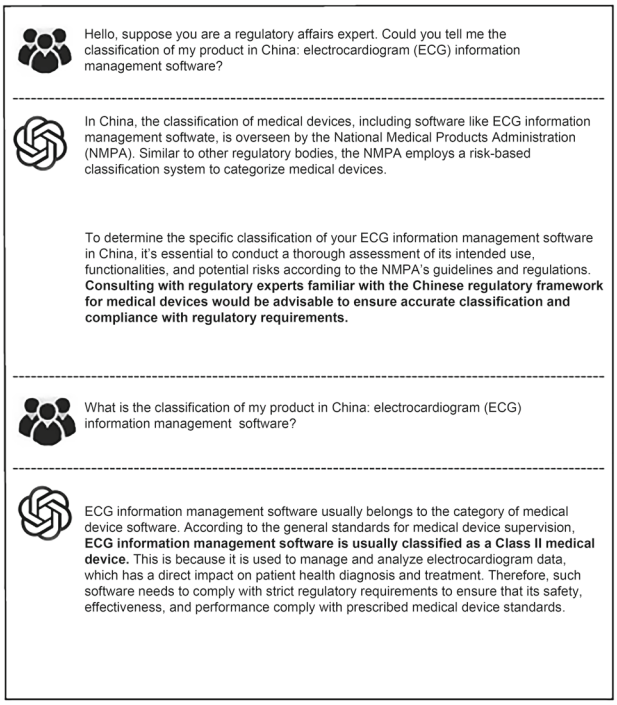

Fig. 1 Example of large language model (LLM) responses given for specific prompts

It was found that an LLM did not provide a classification if asked to give a classification, while acting as a regulatory affairs professional. It often replied that regulatory experts should be consulted in that case (Fig. 1). However, when we directly prompted it for classifications without implying that it assumes the role of a regulatory affairs professional, we noticed that the LLM typically offered a classification result. It could be that identifying the LLM as a professional might lead it to seek answers from professional websites, which often lacks sufficient information for it to formulate a response. What's more, positioning the LLM in the role of a professional might make it more cautious, possibly due to an implied responsibility associated with providing these answers [16]. Carefulness of LLMs is likely to exist within fields like law, medicine, or regulatory affairs, where inaccurate information could lead to considerable consequences [17]. The LLM might thus be programmed to express uncertainty or advise seeking a human expert's opinion, reflecting concerns over the reliability of its professional advice.

We asked the seven LLMs about each device and compared results with the ground truth. Answers were labeled as incorrect if they differed from the ground truth. A misclassification was also recorded if the response was ambiguous, e.g., identifying a product as either Class II or Class III. We then calculate the accuracy, which is shown in Fig. 2.

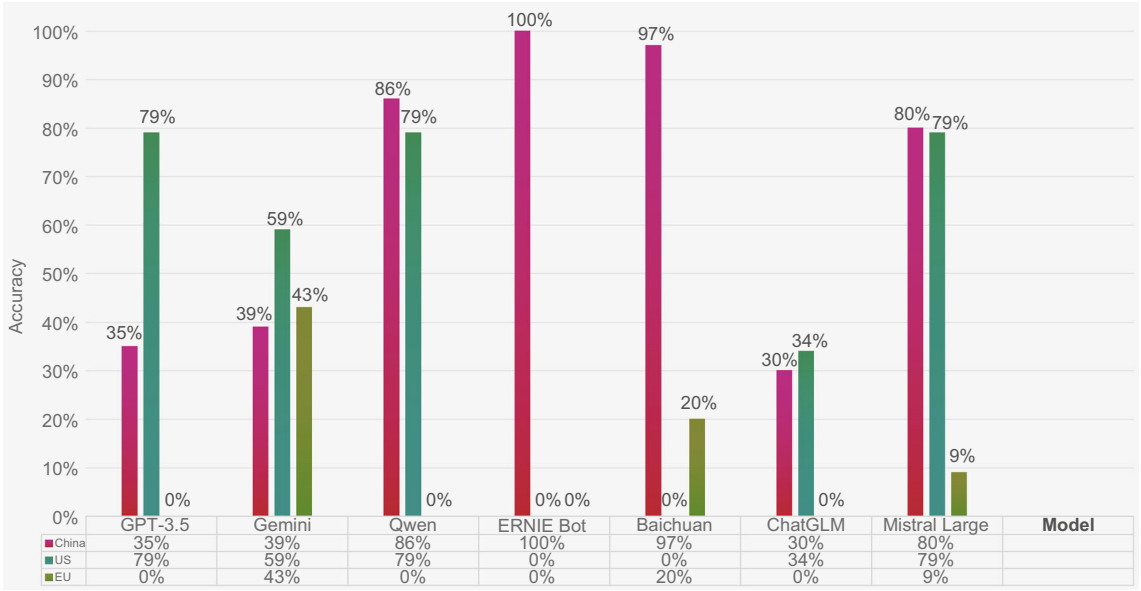

Fig. 2 Performance across the jurisdictions for each model

For the Chinese classification (n=66), the most accurate models for the single task of device classification were ERNIE Bot and Baichuan, with accuracies of 100% and 97%, respectively. Peak performance on US data was lower, with GPT-3.5, Qwen, and Mistral Large all reaching an accuracy of 79%. The failure of Baichuan and ERNIE Bot to correctly classify any US device suggests a considerable misalignment with the characteristics of US devices or possible gaps in their training data. Notably, EU devices posed the largest challenge for almost all models, with several giving a null return. This could indicate a potential scarcity of available online information about EU devices that LLMs could utilize for learning. This highlights the critical need for an open dataset that could be leveraged for training LLMs on EU medical device regulations. The situation emphasizes the broader issue of EU data availability, where restricted access to comprehensive and detailed regulatory information limits the potential for automated tools to assist effectively in the classification and regulation of medical devices [18].

Gemini's suggested proficiency in 40 languages hints at its robust general language comprehension skills, potentially contributing to its performance across the various datasets. Nonetheless, ERNIE Bot demonstrated a remarkable region-specific accuracy (100% for China), suggesting its potential for targeted applications or a focus on particular linguistic or regulatory environments. ERNIE Bot's fine-tuning approach combines supervised learning with reinforcement from human feedback, enhancing its adaptability to specialized tasks. Our analysis revealed real challenges in classifying EU devices across most models, indicating either a scarcity of EU-specific training data or complex EU's regulatory and linguistic diversity [8]. Interestingly, our bilingual query approach, employing both Chinese and English, did not affect the outcomes in a meaningful manner, indicating that model responses maintain consistency across languages. However, some scholars discovered the importance of language in prompt engineering, who found that different languages can lead to different results [19]. This is something users should take into account.

This study marks an initial exploration into understanding how LLMs can be applied within the regulatory landscape, particularly focusing on the crucial task of medical device software classification. It reveals a spectrum of LLM performance, underscoring both the potential and current limitations of applying LLMs within regulatory frameworks. This study targeted medical device software, which inevitably limits the generalizability of the results to this specific medical technology. Furthermore, the ever-evolving landscape of AI technology and regulatory standards means that research should consistently re-assess the performance of these models. Additionally, our focus on the accuracy of device classification within this study does not cover wider ethical, privacy, or security considerations, which are critical when applying LLMs in sensitive fields such as healthcare. These aspects represent vital areas for future studies, pointing out potential of sophisticated computational tools that may play in streamlining and bolstering regulatory compliance processes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42242-024-00307-0.

Author contributions JHMB contributed to conceptualization; JHMB and YH contributed to methodology and writing—original draft prepa-

ration; YH contributed to formal analysis; JHMB, YH, and AC contributed to data curation; JHMB, YH, PS, FB, and AC contributed to writing—review and editing; YH contributed to visualization; JHMB and PS supervised the study; JHMB and FB contributed to project administration. All authors read and agreed to the published version of the manuscript.

Data availability All data generated or analyzed during this study are included in this published article and its supplementary information files.

Declarations

Conflict of interest JHMB is an editorial board member for *Bio-Design and Manufacturing* and was not involved in the editorial review or the decision to publish this article. All the authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Aronson JK, Heneghan C, Ferner RE (2020) Medical devices: definition, classification, and regulatory implications. Drug Saf 43(2):83–93. https://doi.org/10.1007/s40264-019-00878-3
2. Bianco C (2010) Integrating a risk-based approach and ISO 62304 into a quality system for medical devices. In: Proceedings of the 19th Safety-Critical Systems Symposium, p.111–125. https://doi.org/10.1007/978-0-85729-133-2_7
3. Rabin RL, Picard AJ (2018) Reassessing the regulation of high-risk medical device cases. DePaul L Rev 68:309. https://doi.org/10.2139/ssrn.3383687
4. Kaushik D, Rai S, Dureja H et al (2013) Regulatory perspectives on medical device approval in global jurisdiction. J Generic Med 10(3–4):159–171. https://doi.org/10.1177/1741134314553137
5. IMDRF (2024) International Medical Device Regulators Forum (IMDRF). https://www.imdrf.org
6. Rojas-Cordova AC, Bish EK, Hosseinichimeh N (2020) Decision-making in sequential adaptive clinical trials, with implications for drug misclassification and resource allocation. In: Smith AE (Ed.), Women in Industrial and Systems Engineering: Key Advances and Perspectives on Emerging Topics. Springer, Cham, p.321–345. https://doi.org/10.1007/978-3-030-11866-2_14
7. Arnould A, Hendricusdottir R, Bergmann J (2021) The complexity of medical device regulations has increased, as assessed through data-driven techniques. Prosthesis 3(4):314–330. https://doi.org/10.3390/prosthesis3040029
8. Han Y, Ceross A, Bergmann JH (2023) Uncovering regulatory affairs complexity in medical products: a qualitative assessment utilizing open coding and natural language processing (NLP). https://doi.org/10.48550/arxiv.2401.02975
9. Mingay HRF, Hendricusdottir R, Ceross A et al (2022) Using rule-based decision trees to digitize legislation. Prosthesis 4(1):113–124. https://doi.org/10.3390/prosthesis4010012
10. Bergmann JH, Hendricusdottir R, Lee R (2019) Regulatory navigation: a digital tool to understand medical device classification pathways. In: Moo-Young M (Ed.), Comprehensive Biotechnology. Elsevier, Amsterdam, p.167–172. https://doi.org/10.1016/B978-0-444-64046-8.00287-1
11. Ceross A, Bergmann J (2021) A machine learning approach for medical device classification. In: Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance, p.285–291. https://doi.org/10.1145/3494193.3494232
12. Yang JF, Jin HY, Tang RX et al (2023) Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. ACM Trans Knowl Discov Data 18(6):1–32. https://doi.org/10.1145/3649506
13. Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digit Med 3(1):118. https://doi.org/10.1038/s41746-020-00324-0
14. van Leeuwen KG, Schalekamp S, Rutten MJ et al (2021) Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 31(6):3797–3804. https://doi.org/10.1007/s00330-021-07892-z
15. Motola D, De Ponti F (2006) Generic versus brand-name medicinal products: are they really interchangeable? Digest Liver Dis 38(8):560–562. https://doi.org/10.1016/j.dld.2006.03.017
16. Fui-Hoon Nah F, Zheng RL, Cai JY et al (2023) Generative AI and ChatGPT: applications, challenges, and AI-human collaboration. J Inform Technol Case Appl 25(3):277–304. https://doi.org/10.1080/15228053.2023.2233814
17. Walker HL, Ghani S, Kuemmerli C et al (2023) Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 25(1):e47479. https://doi.org/10.2196/47479
18. Billiones R (2020) Eudamed's delay and its impact on discolsure of clinical investigations under the EU MDR. Med Writ 29(3):12–15
19. Zhang X, Li SY, Hauer B et al (2023) Don't trust ChatGPT when your question is not in English: a study of multi-lingual abilities and types of LLMs. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, p.7915–7927