Veeva MedTech

WHITE PAPER

Evaluating Artificial Intelligence in Medtech Regulatory Affairs

Can GenAI ~~take~~ help my Regulatory Affairs job?

Summary

There is increasing interest in generative AI and its applications in medtech regulatory affairs, but organizations need to evaluate this technology with a rigorous lens. In a regulated industry where every document and decision is reviewed and audited, regulatory outputs must be 100% accurate so patient safety is not compromised.

This study compares six leading large language models (LLMs) to human expert performance on several regulatory affairs tasks, ranging from the tactical and discrete to the broad and strategic. While LLMs perform impressively in some respects, they exhibit high variability across both tasks and models. No single model emerged as reliable across all tasks, nor did any model demonstrate an ability to "replace" a human expert on any core regulatory task.

The current state of large language models makes them a useful accessory to human regulatory professionals, but does not support full task outsourcing to algorithmic models. The exercise underscores the importance of human oversight and "human-in-the-loop" design and deployment principles for operational AI in medtech.

Introduction

The release of ChatGPT3 in 2022 introduced the general public to the potential benefits and pitfalls of large language models. LLMs are a type of generative artificial intelligence (GenAI) trained on large datasets that can perform a variety of natural language processing tasks, such as information retrieval or text generation. The models' ability to complete tasks (like creating documents and taking tests) focused attention on GenAI's potential for positive impact, from education to government to business, but also highlighted concerns about unknown risks and the potential for disruption.

Publicly accessible LLMs are prompting medtech companies and individuals to explore how they may improve patient care, medical research, and education.[1] Further, LLMs have the potential to scale up operations by reducing the administrative burden on professionals and shifting resources to value-added tasks.[2] Regulatory affairs (RA) teams, which are being asked to handle an unprecedented volume and complexity of regulatory tasks with the same resources and profitability expectations, can potentially benefit from this technology.

For medical device and diagnostics companies, it is important to ascertain if and how LLMs can benefit regulatory affairs. Can they sift through regulations and provide clear and correct guidance? Can they summarize documents in a comprehensive, efficient, and helpful way? Can they even provide strategic advice?

Methodology and Findings

We looked at six popular, well-regarded LLMs in the market and input three simple exercises, which are representative of different tasks that RA professionals (from specialists to VPs) typically perform. The exercises required information retrieval from regulatory documents, dispensal of technical advice, and provision of strategic guidance.

We provided each LLM with the same initial prompt and input materials and then compared the results from each model on completeness/accuracy, length, and style using a four-point scale:

SCORE	COMPLETENESS/ACCURACY	STYLE	LENGTH
1	Incorrect/Incomplete	Risky	Too short (<150 words) or too long (>850 words)
2	Non-value added	Generic	Short or long
3	Most key-themes referenced	Acceptable	Acceptable
4	All key points referenced	Expert	Executive summary (500 words)

In each exercise, the models' average performance across all categories was benchmarked against expert human advice (defined as an average score of 4 across all categories).

In order to standardize tasks and inputs, we ran all exercises through PromptBros.ai,[3] an AI interaction management system that provides access to the latest commercial GenAI models.

The following models were used:

- Perplexity AI Sonar Huge
- Google Gemini 1.5 Flash
- Anthropic Claude 3.5 Sonnet
- Open AI ChatGPT4
- Meta Llama 3.1 8b
- Mistral AI Le Chat Large

[1] The future landscape of large language models in medicine https://www.nature.com/articles/s43856-023-00370-1

[2] Cegeka boosts Terumo Europe's regulatory efficiency with conversational AI https://www.cegeka.com/en/case-studies/conversational-ai-boosts-terumo-europes-regulatory-efficiency

[3] www.promptbros.ai

EXERCISE 1

Strategic Advice on the AI Act

The Artificial Intelligence Act,[4] published earlier this year, establishes a common regulatory and legal framework for AI within the European Union and will have a significant impact on medtech companies in terms of resources, personnel, risks, and operational complexity.

Model Task

We uploaded 419 pages of the AI Act into the different models and entered the following prompt: *"I want you to act as an expert in GenAI technology and also on regulatory affairs. Please summarize the content of this document focusing on the impact for medtech companies in terms of resources (personnel and budget), compliance risks, and operational complexity."*

Human Comparison

We analyzed a series of articles on the implications of the AI Act for medtech companies from reputable sources (BSI,[5] Clarivate,[6] Hogan Lovells,[7] KPMG,[8] McDermott Will & Avery,[9] TUVSUD[10]) and ranked the underlying key themes by frequency and importance. A total of 14 themes were identified, ranging from risk classification, governance and data management through conformity assessments, device classification, human supervision and training, amongst others.

Model Performance

We sorted the responses from the different models and cross-checked them against the list of key themes for completeness, length, and style.

EXERCISE 1 RESULTS SUMMARY

Model	Perplexity	Mistral	Gemini	OpenAI	Llama	Claude
Completeness	3	3	2	3	2	3
Length	3	3	1	4	2	2
Style	2	2	2	3	3	3
Input	PDF	text	PDF	PDF	PDF	text
Performance vs Human	67%	67%	42%	83%	67%	67%

[4] EU AI Act https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf

[5] Presentation at RAPS Euro Convergence 2024 https://euroconvergence2024.eventscribe.net/fsPopup.asp?PresentationID=1353328&mode=presInfo

[6] 'Regulatory lasagne' and the impact of the European AI Act on medtech https://clarivate.com/blog/regulatory-lasagne-and-the-impact-of-the-european-ai-act-on-medtech/

[7] Implications of the EU AI Act on medtech companies https://www.engage.hoganlovells.com/knowledgeservices/news/implications-of-the-eu-ai-act-on-medtech-companies_1

[8] 10 essential EU AI Act questions businesses need to know https://kpmg.com/ie/en/home/insights/2024/01/eu-artificial-intelligence-act-art-int.html

[9] The Impact of the New EU AI Act on the Medtech and Life Sciences Sector https://www.mwe.com/insights/the-impact-of-the-new-eu-ai-act-on-the-medtech-and-life-sciences-sector/

[10] EU AI Act & ISO IEC IEEE Standards for AI Webinar FAQs https://www.tuvsud.com/en-gb/resource-centre/blogs/uk/testing-and-certification-blog/eu-ai-act-and-iso-iec-ieee-standards-for-ai-and-impact-on-industry

Completeness ranged from 29% to 79% based on the number of key themes referenced, which scored the models between 1 and 3 points on our scale. The length of responses were between 117 and 491 words and we designated response styles as either generic (2) or acceptable (3).

OpenAI, Perplexity, and Claude performed the best in terms of completeness, with OpenAI providing more strategic output in terms of resources and costs and Perplexity providing a very succinct and generic response. Claude's summary was quite complete, however the interface has a word-maximum which limited performance. Mistral and Llama both provided fairly generic summaries, with Llama not being clear if it was medtech-specific and Mistral unable to handle the large document, imposing a word limit. The worst performing model was Gemini, with answers so short and incomplete that we couldn't tell if the document was properly analyzed.

Given the limitations, regulatory teams may choose to leverage LLMs for guided search, but they will still need humans in the loop to correct and augment the summarized outputs.

Exercise 1 Summary

While OpenAI's model scored best overall, it remains relatively weak compared to the human expert assessment of the AI Act. Given the limitations, regulatory teams may choose to leverage LLMs for guided search, but they will still need humans in the loop to correct and augment the summarized outputs. Current LLM versions help this type of task, with correct supervision, but are not a "replacement" for human expertise.

EXERCISE 2

Classifying Medical Devices

Medtech professionals are often required to classify medical devices in order to determine marketing or regulatory pathways.

Model Task

In this exercise, we asked LLMs to review a list of software applications and determine which were medical devices based on the MDCG 2019-11 *Guidance on Qualification and Classification of Software in Regulation for EU MDR and IVDR*,[11] using a list of 13 intended uses for an app or software from the *MHRA Guidance: Medical device stand-alone software including apps (including IVDMDs)*.[12] Models' answers were then challenged by feeding them with the MHRA Guidance examples and querying *"Using examples below from MHRA, would you change any of your answers and why?"*

[11] MDCG 2019-11 Guidance on Qualification and Classification of Software in MDR and IVDR https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2019_11_guidance_qualification_classification_software_en_0.pdf

[12] MHRA Guidance: Medical device stand-alone software including apps (including IVDMDs) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/548090/Medical_device_stand-alone_software_including_apps.pdf

Human Comparison

We sorted and cross-checked the responses from the different models against the MHRA list in order to define completeness/accuracy and style.

Model Performance

EXERCISE 2 RESULTS SUMMARY

Model	Perplexity	Mistral	Gemini	OpenAI	Llama	Claude
Answer 1	13/13	12/13	9/13	13/13	12/13	12/13
Grade 1	4	3	1	4	3	3
Answer 2	12/13	12/13	9/13	13/13	10/13	13/13
Grade 2	2	3	1	4	1	4
Style	3	2	1	4	2	3
Input	PDF	PDF	training dataset	PDF	PDF	PDF
Performance vs Human	75%	67%	25%	100%	50	83%

OpenAI correctly classified all 13 medical device applications and most LLMs performed adequately (12 or more correct answers). In terms of style, LLMs ranged from risky (Gemini), generic (Mistral, Llama), acceptable (Perplexity, Claude) and expert (OpenAI). Gemini did not allow us to upload the classification guidance and used its original training dataset instead, which may explain its poor performance. However, it also confidently stated that the MHRA guidance supported its findings when it expressly did not, a very worrying trait and hence the score of 1 for style.

Interestingly, when we added the MHRA document where the exact wording for each intended use was specifically called out together with its classification by the regulating body, some models revised responses to increase accuracy (Claude), some maintained the original answer (Mistral, Gemini, OpenAI), but others actually changed responses to be less accurate (Perplexity and Llama).

LLMs were more reliable on this tactical task compared to exercise 1, with most models scoring well on completeness and some performing at 100%. However, the volatility of some of the models' responses calls into question repeatability and robustness over time.

Exercise 2 Summary

LLMs were more reliable on this tactical task compared to exercise 1, with most models scoring well on completeness and some performing at 100% (which is what we'd expect from human RA experts). However, the volatility of some of the models' responses calls into question repeatability and robustness over time.

RA professionals may leverage this capability to produce a first draft of simple classifications for further review and ratification by a human RA expert.

EXERCISE 3

Retrieving and Summarizing Information

From looking through records to summarizing technical documents, information retrieval is a fundamental task that regulatory affairs professionals perform day in and day out.

Model Task

This two-part exercise attempted to simulate this common task by asking LLMs to provide an executive summary of a technical document and to highlight specific data points from a graph within it.

First, we fed the models a whitepaper[13] containing a series of facts on the increase of regulatory activities and their potential impact to regulatory affairs operations in medtech. Then we provided the following prompt: *"I want you to act as an expert in GenAI technology and also on regulatory affairs. Please review and provide an executive summary of this white paper, useful for VPs of QARA in medtech companies; use only information from this text."*

Next, we prompted LLMs to retrieve specific data from a chart within the same document with this message: *"This image shows the number of FDA medical guidance documents released over time. The size of the blue bar is the number of guidances and the x axis is the year they were released. The y axis is the number I am looking for. With this information, what was the number of FDA guidances released in 2015?"*

Finally, we challenged the models on the number provided by saying "*Is that response based on what you saw in the graph? It doesn't look aligned to me.*"

Human Comparison

Responses from the different models to both questions were then sorted and cross-checked against a list of 8 high-level themes and 30 details present in this article, as categorized and classified by human RA experts. They were then scored in terms of length and completeness.

Model Performance

EXERCISE 3 RESULTS SUMMARY

Model	Perplexity	Mistral	Gemini	OpenAI	Llama	Claude
Length	3	2	1	2	3	4
Completeness	3	1	1	3	1	4
Image initial	1	2	1	1	1	1
Image challenge	2	2	1	1	2	3
Input	Image	No access	Image	Image	No access	Image
Performance vs. Human	56%	44%	25%	44%	44%	75%

[13] How to Tackle Complex Medtech Regulations with Existing Resources https://www.veeva.com/medtech/resources/complex-regulations/

Three models either did not provide information (Mistral) or worryingly made up a completely new document and title (Gemini, Llama). The other three models provided quite complete summaries covering at least 7 of the 8 key themes and 16 of the 30 details in the text. The best performing model was Claude (8 key themes, 16 details), followed by OpenAI (7 key themes, 22 details), which was downgraded to a score of 3 as it provided a made-up fact not present in the document.

When asked to retrieve information from the graph, all models performed poorly to start (scores of 1 or 2), with answers hinting that they were nothing but a guess. While the graph clearly showed that the correct number of FDA guidances released in 2015 was 7, the models' initial guesses ranged between 10 and 34.

However, when challenged about initial guesses, some models updated them to approach the correct answer, but only one could provide an accurate approximation (Claude). Other models provided cautious answers although still wrong. Mistral expressly stated that it had no visual access and it refused to provide an answer, which was reassuring. In contrast, Llama also did not allow for image access or upload but provided the answer that mostly deviated from the target (34) as a matter of fact.

LLMs' overall performance in exercise 3 was uneven, including issues with "hallucinations", and their inability to correctly complete the image task was troubling.

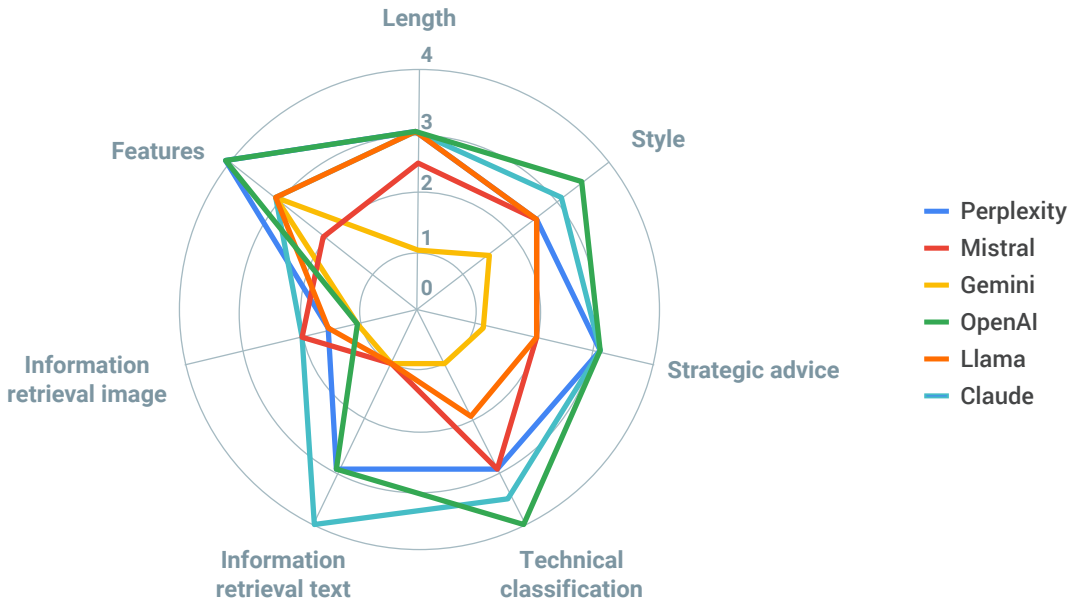
Exercise 3 Summary

LLMs' overall performance in exercise 3 was uneven, including issues with "hallucinations", and their inability to correctly complete the image task was troubling. Similar to exercise 1, RA professionals may employ LLMs to quickly summarize long-form text, but they'll need to seriously scrutinize the output.

Discussion and Conclusion

On average, LLM performance ranged between 25% and 79% across all exercises when benchmarked against human experts (100%), highlighting that regulatory affairs professionals may only want to use them as a "sparring partner" when analyzing and applying regulatory guidance.

Although useful for some text-based activities, like executive summaries of technical documents, LLMs did not provide much strategic advice and often missed key points that could have a significant impact on medtech resources. For example, most models forgot to consider the impact of extra conformity assessments and issues with Notified Body capacity as part of the strategic advice provided regarding the EU AI Act (Exercise 1). Some LLMs were useful at providing technical input when fed guidance classification (see Exercise 2), but this behavior was erratic. On average, LLM performance ranged between 25% and 79% across all exercises when benchmarked against human experts (100%), highlighting that regulatory affairs professionals may only want to use them as a "sparring partner" when analyzing and applying regulatory guidance.



Not surprisingly, LLMs were generally better at performing text-based tasks rather than image retrieval, with exercise 3 being the worst performing across all models (average score 1.92 vs 2.58 for Exercise 1 and 2.67 for Exercise 2). Given that these models were designed to use artificial neural networks to process and generate large-scale text data in a way where shape matters more than content, regulatory affairs professionals should be mindful of their limitations outside those parameters.

A key finding from this report is that the performance varied greatly between models within each exercise and with no clear pattern across exercises.

For example, some models performed the best on Exercises 1 and 2, but then struggled in Exercise 3, where they made up data and quotes. Others were consistently poor. Interestingly, models had mixed summarizing performance across Exercise 1 and 3, with some LLMs providing a fairly generic and short strategic assessment of the impact of the AI Act, but having a much better performance on the white paper exercise.

It is worrying that, when challenged in exercise 2 and 3, some models revised correct answers to wrong ones. Counter-intuitively, they got worse when further guidance was provided. Alarmingly, some models made up completely new documents when fed the whitepaper in Exercise 3.

This is certainly an area worth exploring by medtech companies, provided they set clear business objectives, take appropriate risk controls, and develop technical platform and data strategies to provision model data.

The study is not comprehensive, as this is a rapidly developing area of artificial intelligence with evolving models and data sets. As we increase our understanding of LLMs and how they get to certain answers, prompt engineering will also surely improve. This is certainly an area worth exploring by medtech companies, provided they set clear business objectives, take appropriate risk controls, and develop technical platform and data strategies to provision model data.[14] Setting up an AI strategy that delivers value is an important first step in this direction. It should outline clear objectives, timelines and performance metrics to track progress and measure value.

Lastly, it's important for regulatory teams to be aware of key technical and data privacy considerations in real-world medtech implementations of operational AI. For these exercises, we used current/late versions of commercial LLMs, but due to IP and data privacy considerations, industry generally bars the use of public models from behind corporate firewalls. Development of in-house LLMs have limited potential for further model training beyond their internal data. We should therefore expect real-world industry implementations on any RA tasks to be significantly worse than the rapidly evolving public models.

Overall, LLMs fell short compared to expert human performance. This underscores the importance of deploying AI thoughtfully using a best practice technology stack and controls framework, supported by partners with the right technical, business, and regulatory competencies. To our understanding this is the first study of GenAI and LLMs within the context of regulatory affairs in medtech, but we expect to see more research on how this technology can support RA teams in the future. In the meantime, we can attempt to answer the tongue-in-cheek questions that inspired this work…

Will GenAI help my regulatory affairs job? Probably yes.
Will GenAI take my regulatory affairs job? Probably not.

[14] AI is a Starting Point, Not a Magic Wand https://www.veeva.com/medtech/resources/ai-is-a-starting-point-not-a-magic-wand-2/

Author

Diogo Geraldes
Director, Regulatory Strategy, Veeva MedTech
diogo.geraldes@veeva.com

Diogo is a Director for Regulatory Strategy at Veeva Systems, specializing in the EU medtech market. Previously, Diogo was the UKCA certification manager and a principal technical assessor at DNV, ensuring the safety and compliance of medical devices for the EU and UK markets. Diogo's career also includes work as a designer at Stryker/Stanmore Implants, where he contributed to the development of over 250 patient-specific implants, particularly focusing on complex cases in limb salvage and pediatric oncology. With a background in biomechanics, Diogo earned his PhD at Imperial College London, where he focused on computational modelling of bone adaptation in the femur. He followed this with a post-doctorate at Imperial, where he developed and patented a novel glenoid implant. Diogo combines academic rigor with design, research, modelling, and experience in regulatory affairs.