Large Language Models and Regulatory-Grade GLP-1 Content: Challenges and the Need for Accurate Message Measurement

Authors: Anna Ruddell, Roma English Owen, Kirsty Mursec, Will Staples, Josh Eadsforth, Elizabeth Fairley
Published Online: 27 April 2026

Glucagon-like peptide-1 receptor agonists (GLP-1s) have rapidly emerged as a transformative therapeutic class in metabolic disease management, necessitating stringent regulatory oversight and precise, up-to-date information. This study evaluates the accuracy of large language models (LLMs), specifically OpenAI's ChatGPT and Anthropic's Claude, in generating responses compared to U.S. Food and Drug Administration (FDA) GLP-1 regulatory content. Additionally, it examines the reliability and validity of LLM-based scoring methods, using ChatGPT and Copilot, to assess outputs against a standardized rubric. Analyses reveal inconsistencies in LLM-generated content, including omissions and incorrect or misleading information, highlighting current limitations of LLMs for autonomous content generation amid the dynamic landscape of GLP-1s. Variability in scoring between evaluating LLMs further underscores subjectivity inherent in automated assessments. These findings highlight the need for quality-controlled, objective metrics such as Message Resonance Score™. Such tools are essential for measuring the alignment of messaging to real-life captured healthcare data, ultimately supporting safer and more effective integration of GLP-1 therapies in clinical practice.

Introduction

Large language models (LLMs) such as OpenAI's ChatGPT and Anthropic's Claude have demonstrated capabilities in natural language understanding and generation.[4] Their ability to synthesise complex information from authoritative sources holds promise for regulatory and healthcare domains, where accurate and evidence-based communication is critical.[5,6] Despite advances in LLM performance, challenges remain in ensuring the factual accuracy, completeness, and reliability.[7]

Glucagon-like peptide-1 receptor agonists (GLP-1s) have recently gained significant attention due to their expanding role in managing metabolic diseases, alongside emerging safety considerations. The U.S. Food and Drug Association (FDA) have recently approved a second oral GLP-1, marking the constantly evolving GLP-1 landscape.[15] This underscores the need for reliable synthesis and evaluation of regulatory information related to GLP-1s.

In this study, we compare the responses generated by ChatGPT and Claude to questions based on an FDA press release concerning GLP-1s. We further analyze the scoring patterns of two LLM-based evaluators (ChatGPT and Copilot) to understand differences in evaluation strictness and consistency. We hypothesize that LLM-generated outputs will diverge from established regulatory truths, and that evaluations conducted by LLMs will exhibit inconsistencies, even when applying the same grading criteria.

Methodology

Two LLMs, GPT-5.3 ChatGPT[12] and Sonnet 4.6 Claude[1] were prompted to answer two specific GLP-1 questions based on an FDA press release titled "FDA Approves First New Molecular Entity Under National Priority Review Voucher Program".[15] The questions were:

1. What is the significance of the FDA's recent approval mentioned in the press release? Which drug was approved under this program?
2. Explain the purpose of the National Priority Review Voucher Program as described in the FDA's announcement, and how the recent drug approval fits within this program.

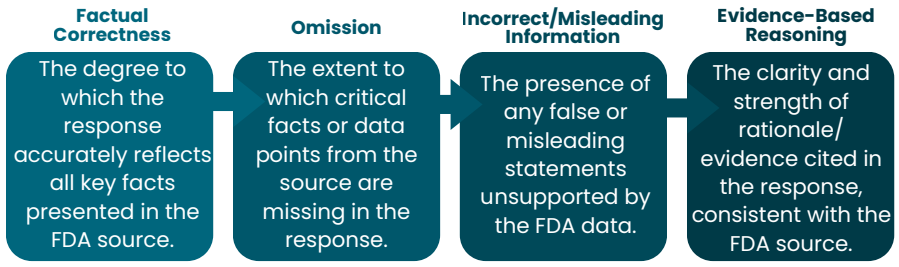
The LLM prompt used: "Using [FDA source hyperlink] [Question X], was asked to both ChatGPT and Claude.

Two LLMs, GPT-5.3 ChatGPT[12] and Copilot[10], independently evaluated the responses using a predefined rubric (Figure 1) on four dimensions (Figure 2). Each dimension was scored on a 6-point scale (0–5) of higher scores indicating better performance.

Figure 1. *Simplification of the Full Predefined Rubric for Evalutating LLM-Generated Content*

Dimension	Definition	Scoring Parameter (0)	Scoring Parameter (5)
Factual Correctness	Does the LLM response include all key facts accurately as per FDA content?	0 = Completely incorrect or irrelevant information	5 = Fully accuracte and comprehensive, reflecting all key facts from FDA content
Omission	Are any critical facts or data points from the FDA source missing in the LLM response?	0 = Major omissions of essential information that significantly affect understanding	5 = No omissions; all critical information is included
Incorrect/Misleading Info	Does the response contain any false or misleading statements not supported by FDA data?	0 = Numerous false or misleading statements present	5 = Completely free of false or misleading information
Evidence-based Reasoning	Does the LLM provide rationale or cite evidence with FDA information?	0 = No reasoning or evidence provided	5 = Clear, well-supported, and explicit evidence-based reasoning fully aligned with FDA data

Figure 2. *Shows the Criteria of the 4 Dimensions Each Scored (0-5)*



The LLM evaluation prompt used: "Compare [LLM output] to [FDA source hyperlink] using [predefined rubric]".

Results

The scores were averaged (N = 4) to evaluate the accuracy of LLMs (ChatGPT and Claude) compared to the GLP-1 FDA source (Figure 3) and to compare the scores assigned by the LLMs (ChatGPT and Copilot), as seen in Figure 4. This evaluation used the full predefined rubric, which is not shown here; a simplified version appears in Figure 1.

Figure 3. *Mean Accuracy Scores of ChatGPT and Claude Outputs Compared to the GLP-1 FDA Source*

Dimension	ChatGPT Accuracy (0-5)	Claude Accuracy (0-5)
Factual Correctness	4.25	4.50
Omission	2.75	4.25
Incorrect/Misleading Info	4.50	4.50
Evidence-based Reasoning	4.00	5.00

Figure 4. *Comparison of Mean Evaluation Scores for ChatGPT and Copilot Using the Same Scoring Rubric*

Dimension	ChatGPT Evaluation (0-5)	Copilot Evaluation (0-5)
Factual Correctness	4.25	4.50
Omission	3.00	4.00
Incorrect/Misleading Info	4.25	4.75
Evidence-based Reasoning	4.50	4.50

Figure 5. *Graph to Show LLM Output Accuracy Compared to FDA GLP-1 Source*

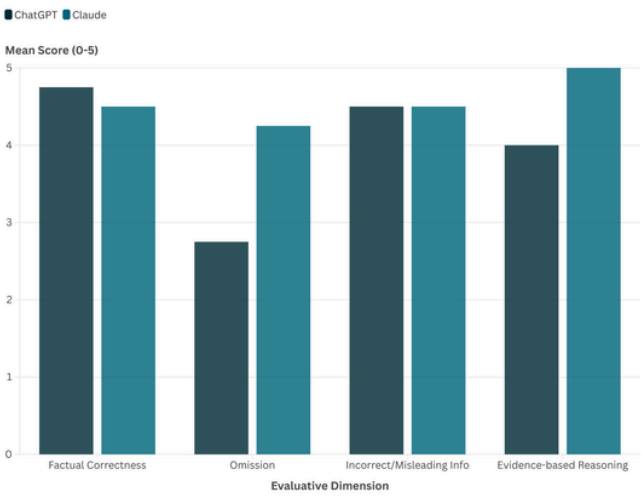
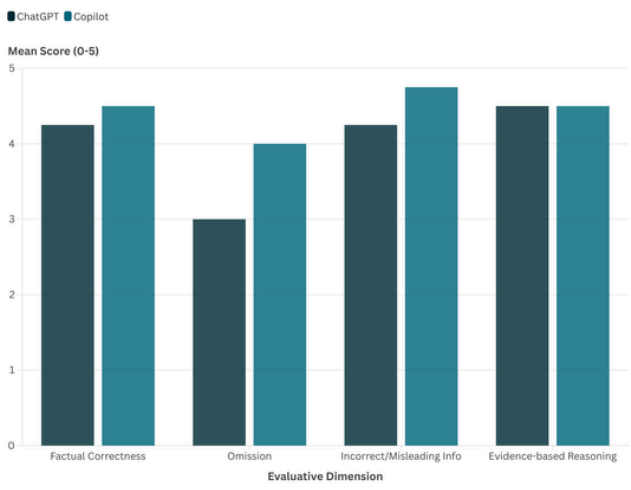


Figure 5 shows that LLM outputs vary in accuracy scores and do not reach regulatory-level accuracy. Claude's outputs were generally rated higher in omission and evidence-based reasoning, and contained less incorrect or misleading information. ChatGPT's outputs were seen as slightly more factually correct.

Figure 6. *Graph to Show Scoring Patterns of Evaluating LLMs: ChatGPT vs. Copilot*



LLM (ChatGPT and Copilot) scoring patterns differed across most dimensions, as seen in Figure 6. Copilot had more lenient scoring and generally assigned higher scores across most dimensions compared to ChatGPT. ChatGPT provided more critical assessments, particularly evident in the lower mean score for omission.

Results

A focused qualitative comparison was conducted on the completeness and accuracy of information generated by ChatGPT and Claude relative to the FDA source regarding the drug approved under the GLP-1 receptor agonist program.

The FDA source states the GLP-1 is approved with diet and exercise to reduce and maintain weight loss in adults with obesity or overweight plus related conditions, as seen in Figure 7.

Figure 7. *FDA Source - Press Release Description of GLP-1 Approval and Indication*

> Foundayo is approved for use in combination with a reduced-calorie diet and increased physical activity to reduce excess body weight and maintain weight reduction long term in adults with obesity or adults with overweight in the presence of at least one weight-related comorbid condition.

ChatGPT's output (Figure 8) succinctly identified the approved drug and correctly listed critical attributes, including its status as a new molecular entity, its oral administration route, and its indication for chronic weight management in overweight or obese patients.

Figure 8. *ChatGPT-Generated Summary*

> **Drug approved under the program**
>
> - The approved drug is **Foundayo (orforglipron)**.
> - It is:
>   - A new molecular entity (NME)
>   - An oral (non-injectable) GLP-1 receptor agonist
>   - Approved for chronic weight management (overweight/obesity)

In contrast, Claude's output (Figure 9) provided a more detailed narrative, including the developer's name, the approval date, and additional contextual information about the drug's mechanism and clinical use.

Figure 9. *Claude-Generated Summary*

> **The Drug: Foundayo (orforglipron)**
>
> The FDA approved **Foundayo (orforglipron)**, developed by **Eli Lilly and Company**, on April 1, 2026. It is a once-daily oral tablet and a GLP-1 receptor agonist approved for long-term weight management in adults with obesity or overweight with at least one weight-related comorbid condition, used alongside a reduced-calorie diet and increased physical activity. fda

The results reveal notable differences both in the quality of the LLM-generated outputs and in the scoring patterns of the evaluating LLMs.

The evaluation of LLM outputs against the FDA GLP-1 source showed that both ChatGPT and Claude demonstrated varying levels of accuracy across the key dimensions of factual correctness, omission, incorrect/misleading information, and evidence-based reasoning. Claude generally scored higher in evidence-based reasoning and completeness, while ChatGPT was slightly more accurate in factual correctness.

When scored using the same rubric, the two evaluating LLMs (ChatGPT and Copilot) showed distinct scoring patterns. Copilot was more lenient by generally assigning higher scores across all metrics and questions, whereas ChatGPT provided more critical and stringent assessments. This contrast highlights the variability in evaluation strictness between models, when provided with the same scoring framework.

Discussion

This study demonstrates significant challenges in relying solely on large language models (LLMs) to generate regulatory-grade content in the GLP-1 therapeutic area. The evaluation revealed notable differences in both the quality of LLM-generated outputs and the scoring patterns of the evaluating LLMs. Compared to competitors, this work uniquely assesses LLM accuracy against authoritative FDA GLP-1 sources and evaluates scoring consistency using a standardized rubric. While ChatGPT provided more critical and stringent assessments, Copilot was more lenient, highlighting variability in evaluation strictness that could impact clinical and regulatory decision-making.[4,2]

The clinical value of this differentiation is substantial. GLP-1 therapies are rapidly evolving treatment paradigms for metabolic diseases, with complex safety, efficacy, and regulatory considerations.[8,11] In this context, precision and trustworthiness of information are paramount, as even minor inaccuracies or omissions can have significant consequences.[14] The findings underscore the need for quality-controlled, objective metrics, such as Talking Medicines DrugVoice Message Resonance Score™, to provide consistent, accurate, and clinically meaningful assessments of LLM-generated regulatory content. In addition to objective metrics, Talking Medicines Advanced Data Science and AI platform DrugVoice enables the qualitative discovery of key topics by; the curation, structuring and human in the loop (HITL) quality control of real-life data prior to of LLMs. Such tools will be critical to support clinicians, regulators, and other stakeholders in confidently measuring the alignment of approved messaging to real-life captured data in the dynamic GLP-1 landscape.[13]

This study has several limitations that should be acknowledged. First, the evaluation focused on only two LLMs, ChatGPT and Claude, which may not represent the full spectrum of available or emerging models.[3] Second, the assessment was limited to GLP-1 regulatory content based on FDA sources, and findings may not generalize to other therapeutic areas or regulatory environments. Third, while the rubric used for scoring was comprehensive, the inherent subjectivity in interpreting complex regulatory language may have influenced both LLM outputs and scoring variability.[9] Finally, the study did not assess the impact of iterative prompt engineering or fine-tuning, which could potentially improve LLM accuracy and evaluation consistency.[16] Future research should explore these factors to better understand and enhance the role of LLMs in regulatory science.

Conclusion

This study highlights critical limitations in the application of LLMs, specifically ChatGPT and Claude, for generating regulatory-grade content based on FDA sources within the GLP-1 therapeutic area. The observed discrepancies in factual accuracy and omissions underscore that LLM-generated outputs remain inherently fallible and are not yet suitable replacements for rigorous regulatory standards in the rapidly evolving GLP-1 landscape. Furthermore, the divergent scoring behaviors between evaluating LLMs, Copilots's leniency versus ChatGPT's critical rigor, reveal the current unreliability and subjectivity of LLM evaluation methods.

Given the complexity and clinical significance of GLP-1 therapies, where precision and trustworthiness are paramount, these findings emphasize the urgent need for quality-controlled, objective metrics such as Message Resonance Score™.[8] Talking Medicines unlocks value by transforming unstructured data into actionable intelligence and such tools are essential for measuring the alignment of messaging to real-life captured healthcare data, ultimately supporting safer and more effective integration of GLP-1 therapies in clinical practice.

References

1. Anthropic. Sonnet 4.6 Claude [large language model]. 2026 [cited 2026 Apr 13]. Available from: https://www.anthropic.com/
2. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. p. 610–23. Available from: https://doi.org/10.1145/3442188.3445922
3. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv [Preprint]. 2021. Available from: https://arxiv.org/abs/2108.07258
4. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901. Available from: https://arxiv.org/abs/2005.14165
5. Cossu M, Ciucci D. Leveraging large language models for abstractive summarization of Italian legal news. Artif Intell Law. 2025. Available from: https://doi.org/10.1007/s10506-025-09431-3
6. Kientz JA, Schueller SM, Mohr DC. Mobile behavioral health coaching as a preventative intervention for occupational public health: Retrospective longitudinal study. JMIR Form Res. 2023;7(1):e456784. Available from: https://formative.jmir.org/2023/1/e456784
7. Lin Z, Wang Y. Survey of hallucination in natural language generation. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2023. p. 1–12. Available from: https://doi.org/10.1145/3571730.5
8. Longoni C, Bonezzi A, Morewedge CK. What we talk about when we talk about trust: Theory of trust for AI in healthcare. Patterns. 2020;1(6):100089. Available from: https://doi.org/10.1016/j.patter.2020.100089
9. Marcus G, Davis E. GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review. 2020 Aug 22. Available from: https://www.technologyreview.com/2020/08/22/1007539/gpt-3-openai-language-generator-artificial-intelligence-ai-opinion/
10. Microsoft. Copilot [large language model]. 2026 [cited 2026 Apr 20]. Available from: https://copilot.microsoft.com/
11. Nauck MA, Meier JJ. Incretin hormones: Their role in health and disease. Diabetes Obes Metab. 2019;21(S1):5–21. Available from: https://doi.org/10.1111/dom.13619
12. OpenAI. GPT-5.3 ChatGPT [large language model]. 2026 [cited 2026 Apr 13]. Available from: https://chat.openai.com/
13. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31–38. Available from: https://doi.org/10.1038/s41591-021-01614-0
14. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. Available from: https://doi.org/10.1038/s41591-018-0300-7
15. U.S. Food and Drug Administration. FDA approves first new molecular entity under national priority voucher program [Internet]. 2026 Apr 1 [cited 2026 May 21]. Available from: https://www.fda.gov/news-events/press-announcements/fda-approves-first-new-molecular-entity-under-national-priority-voucher-program
16. Zhao W, Wang Y, Yatskar M, Ordonez V, Chang KW. Calibrate before use: Improving few-shot performance of language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. p. 8659–71. Available from: https://doi.org/10.18653/v1/2023.acl-long.615