










# Comparative Analysis of Generative Pre-Trained Transformer Models in Oncogene-Driven Non–Small Cell Lung Cancer: Introducing the Generative Artificial Intelligence Performance Score

Zacharie Hamilton, BA<sup>1</sup> ; Aseem Aseem, MD<sup>1</sup> ; Zhengjia Chen, PhD<sup>1</sup> ; Noor Naffakh, PharmD, MS<sup>1</sup> ; Natalie M. Reizine, MD<sup>1</sup> ; Frank Weinberg, MD, PhD<sup>1</sup> ; Shikha Jain, MD<sup>1</sup> ; Larry G. Kessler, ScD<sup>2</sup>; Vijayakrishna K. Gadi, MD, PhD<sup>1</sup> ; Christopher Bun, PhD<sup>3</sup>; and Ryan H. Nguyen, DO<sup>1</sup> 

DOI <https://doi.org/10.1200/CCI.24.00123>

## ABSTRACT

**PURPOSE** Precision oncology in non–small cell lung cancer (NSCLC) relies on biomarker testing for clinical decision making. Despite its importance, challenges like the lack of genomic oncology training, nonstandardized biomarker reporting, and a rapidly evolving treatment landscape hinder its practice. Generative artificial intelligence (AI), such as ChatGPT, offers promise for enhancing clinical decision support. Effective performance metrics are crucial to evaluate these models' accuracy and their propensity for producing incorrect or hallucinated information. We assessed various ChatGPT versions' ability to generate accurate next-generation sequencing reports and treatment recommendations for NSCLC, using a novel Generative AI Performance Score (G-PS), which considers accuracy, relevancy, and hallucinations.

**METHODS** We queried ChatGPT versions for first–line NSCLC treatment recommendations with an Food and Drug Administration–approved targeted therapy, using a zero-shot prompt approach for eight oncogenes. Responses were assessed against National Comprehensive Cancer Network (NCCN) guidelines for accuracy, relevance, and hallucinations, with G-PS calculating scores from –1 (all hallucinations) to 1 (fully NCCN-compliant recommendations). G-PS was designed as a composite measure with a base score for correct recommendations (weighted for preferred treatments) and a penalty for hallucinations.

**RESULTS** Analyzing 160 responses, generative pre-trained transformer (GPT)–4 outperformed GPT–3.5, showing higher base score (90% v 60%;  $P < .01$ ) and fewer hallucinations (34% v 53%;  $P < .01$ ). GPT–4's overall G-PS was significantly higher (0.34 v –0.15;  $P < .01$ ), indicating superior performance.

**CONCLUSION** This study highlights the rapid improvement of generative AI in matching treatment recommendations with biomarkers in precision oncology. Although the rate of hallucinations improved in the GPT–4 model, future generative AI use in clinical care requires high levels of accuracy with minimal to no room for hallucinations. The GP-S represents a novel metric quantifying generative AI utility in health care compared with national guidelines, with potential adaptation beyond precision oncology.

## ACCOMPANYING CONTENT

 [Data Supplement](#)

Accepted October 4, 2024  
Published December 11, 2024

JCO Clin Cancer Inform  
8:e2400123  
© 2024 by American Society of  
Clinical Oncology

Creative Commons Attribution  
Non-Commercial No Derivatives  
4.0 License

## INTRODUCTION

Lung cancer is the leading cause of cancer mortality worldwide, with non–small cell lung cancer (NSCLC) being the most common type.<sup>1</sup> *EGFR* mutations and *ALK* rearrangements are well-established drivers for NSCLC tumorigenesis and serve as predictive biomarkers for targeted

drugs.<sup>1</sup> The Food and Drug Administration (FDA) has approved numerous drugs either alone or in combination for biomarker-directed therapy for advanced NSCLC (aNSCLC).<sup>2</sup> As actionable molecular targets in aNSCLC increase, biomarker testing of lung cancer biopsy specimens via next-generation sequencing (NGS) has become routine clinical practice.<sup>3,4</sup> NGS reports are generated to highlight the

## CONTEXT

### Key Objective

This study evaluated the performance of generative pre-trained transformer (GPT)-3.5 and GPT-4 models in generating accurate next-generation sequencing reports and treatment recommendations for non-small cell lung cancer (NSCLC). The study introduces a novel metric, the Generative artificial intelligence Performance Score (G-PS), which considers accuracy, relevancy, and hallucinations.

### Knowledge Generated

GPT-4 outperformed GPT-3.5 in generating accurate treatment recommendations for NSCLC, with a higher base score (90% v 60%;  $P < .01$ ) and fewer hallucinations (34% v 53%;  $P < .01$ ). The overall G-PS was significantly higher for GPT-4 (0.34 v -0.15;  $P < .01$ ).

### Relevance

The authors present a score that can be used to evaluate the quality of large language models—with this, improvements between versions can be measured, and it may also prove to be a valuable tool for cross-comparisons.

genetic alterations and subsequent treatment and prognostic implications to assist patients and oncologists in clinical decision making, prognostication, and treatment selection.<sup>1,3</sup>

The landscape of oncogenic driver mutations and their associated targeted therapies is dynamic and rapidly evolving, posing significant challenges in effectively integrating large amounts of genomic data into routine cancer care.<sup>4-6</sup> Advances in NSCLC treatments including targeted and immune checkpoint treatment have led to significant clinical benefits in the past decade.<sup>7</sup> However, many patients are not matched with appropriate personalized treatments because of clinical practice gaps. A real-world database study found that among patients diagnosed with aNSCLC after 2017, only 77% received at least one biomarker test and 49% received NGS at any point, with Black patients less likely to receive NGS testing before first-line (1L) therapy and at any-given time for aNSCLC. In a separate claims-based study, among those patients who do receive testing, nearly 30% of patients did not receive appropriate targeted therapy.<sup>8</sup> Suspected reasons for these disparities include lack of standardized genomic reporting, awareness of targeted treatment options and/or guidance, institutional access to molecular tumor boards, and social determinants of health affecting access to medications.<sup>5,6,9</sup> Without standardized genomic education in oncology training, oncologists face a rapidly growing amount of biomarker testing to integrate into cancer care in NSCLC and beyond. In 2023 alone, 12 treatments were approved by the FDA for unique biomarker indications and six biomarker and indication-specific treatments were added to the National Comprehensive Cancer Network (NCCN) guidelines.<sup>10</sup>

ChatGPT is a large language model (LLM) artificial intelligence (AI) chatbot platform created by OpenAI that generates text in response to user prompts. As of May 2024,

OpenAI has released two publicly available generative pre-trained transformer (GPT) models, GPT-3.5 and its successor GPT-4, publicly released in November 2022 and March 2023, respectively. Each model was independently trained with publicly available online text sources, with both models trained on data up until September 2021 at the time of our analysis.<sup>11</sup> While GPT-3.5 was built with 175 billion parameters, GPT-4 was built with significantly more (the exact number has not been publicly detailed by OpenAI), allowing for better understanding and generation capabilities. In LLM, parameters refer to the weights and biases in the neural network that determine how inputs (data) are transformed into outputs (predictions or text).<sup>12</sup> The platform interacts with users conversationally and has potential uses across the health care spectrum, including in clinical decision support.<sup>13,14</sup> ChatGPT has demonstrated impressive clinical aptitude on USMLE and other board-style examination questions<sup>15-17</sup> and in achieving correct diagnoses in clinical vignettes.<sup>18</sup> In oncology, ChatGPT accurately answered commonly posed lung cancer questions and GPT-4 performed as well as or better than task-specific models in classifying breast cancer pathology.<sup>19</sup> Furthermore, an analysis of four platforms (including ChatGPT) found that these chatbots could generate high-quality responses that minimized misinformation and had moderate understandability.<sup>20</sup>

One of the limitations of generative AI and the broader field of natural language generation (NLG) is the tendency of these models to generate incorrect or nonsensical responses, commonly referred to as hallucinations.<sup>21</sup> As models are prompted with tasks on which they are not specifically trained (zero-shot prompting), their performance can be unpredictable.<sup>22</sup> Hallucinations are of concern as they hinder optimal model performance and raise concerns for clinical use through spread of misinformation that may result in mistreatment.<sup>13,20,23</sup> While hallucinations and related

evaluation metrics are well-established areas of NLG research, assessments of hallucinations in clinical contexts are comparatively in their infancy.<sup>21,24</sup> This is clear in the diversity in performance assessment approaches in previous studies.<sup>13-20,25-28</sup> When assessing the ability of LLMs in classification tasks, most studies report pure accuracy percentages or manually evaluated Likert scales (ie, five-point rating scale) from human raters. More sophisticated analyses leverage sophisticated metrics such as BLEU and ROUGE-1 to evaluate the quality of machine translation compared with human references.<sup>29,30</sup> In many cases, hallucinations are not explicitly included in the analysis. While ordinal metrics provide a framework to quantify perceived overall correctness, they remain limited in their ability to be extrapolated for more rigorous quantitative analysis and in their current applications in clinical generative AI research which lack standardization.<sup>31</sup> Furthermore, quantitative metrics like BLEU and ROUGE-1 do not correlate well with human judgment regarding hallucinations.<sup>21</sup>

To address the need for a quantitative performance-based assessment metric that factors the negative effects of hallucinations in clinically oriented generative AI for oncology, we designed a cross-sectional comparative study to (1) compare the performance of GPT-3.5 and GPT-4 and (2) pilot the Generative AI Performance Score (G-PS), an LLM performance assessment tool we developed to more effectively account for both accuracy and hallucinations.

## METHODS

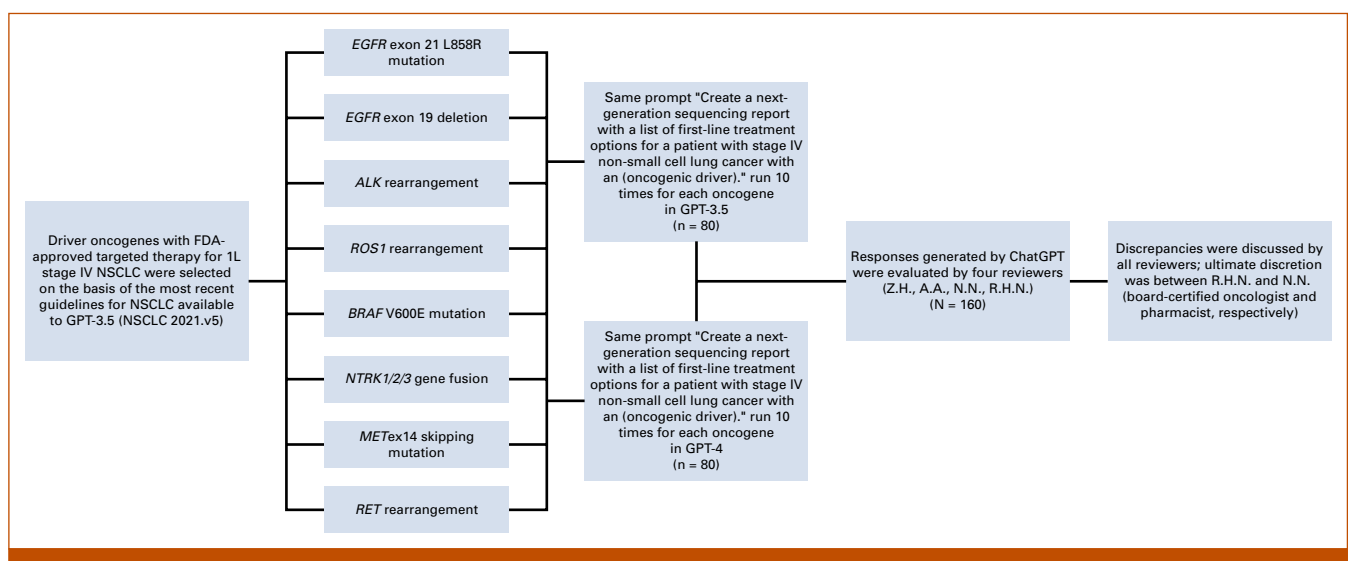
This study used a cross-sectional comparative design to assess the performance of the ChatGPT model in generating text in the style of an NGS report listing 1L treatment options for patients with stage IV NSCLC. Eight driver oncogenes with FDA-approved targeted therapy for 1L stage IV NSCLC

(Fig 1) were selected on the basis of the most recent NCCN guidelines for NSCLC available to the ChatGPT 3.5 and 4 database (NSCLC 2021.v5).<sup>32</sup>

For data collection, we used the same zero-shot prompt, "Create a next-generation sequencing report with a list of first-line treatment options for a patient with stage IV non-small cell lung cancer with an (oncogenic driver)." The bracketed text represented one of the driver mutations (eg, *ALK* mutation). This prompt was run on both the GPT-3.5 (accessed January 29-30, 2023) and GPT-4 (accessed June 30, 2023-July 2, 2023) web-interface versions of ChatGPT (OpenAI). We executed the prompt 10 times for each of the eight selected oncogenes per model, all in a new chat to ensure that each evaluation was independent, unbiased, and free from contextual errors carried over from previous interactions. Responses generated by ChatGPT were evaluated by four reviewers (Z.H., A.A., N.N., R.H.N.). The Data Supplement (Table S1) details the inclusion criteria for a treatment recommendation.

Recommendations were incorporated into the assessment only if they were explicitly indicated by ChatGPT for possible use in a 1L setting, irrespective of any additional context or qualifications provided by the model regarding a treatment's suitability in the 1L setting. This standardized evaluation approach was designed to minimize subjectivity and ensure consistent analysis across all generated responses.

To evaluate model performance, we devised the G-PS (Eq 1). The G-PS score consists of a base score for correctly listed NCCN treatments minus a Hallucination Penalty for inappropriate recommendations. The G-PS uses multiple parameters in both the base score and Hallucination Penalty to provide flexibility in future use cases, particularly when hallucinations are deemed more, or less, permissible in a



**FIG 1.** Flowchart summarizing methodology and flow of evaluation. 1L, first-line; FDA, Food and Drug Administration; GPT, generative pre-trained transformer; NSCLC, non-small cell lung cancer.

particular context. G-PS scoring was designed with a range of  $[-1, 1]$ , from a maximum score of 1 (all treatments listed and no hallucinations) to a minimum score of  $-1$  (all recommendations hallucinated).

The base score accounts for model accuracy in generating recommended treatments and is the weighted sum for preferred (parameter  $x$ ) and other recommended (parameter  $y$ ), set to  $x = 0.75$ ,  $y = 0.25$ , respectively. These parameters are based on the clinical assessment of the authors for the relative importance of using NCCN-preferred treatment recommendations in 1L advanced aNSCLC. The Hallucination Penalty is derived from a logistic regression formula, where  $h$  equals the number of incorrectly listed treatments (ie, hallucinations). The  $a$  parameter acts as a scaling factor, controlling how sensitive the G-PS is to hallucinations. The parameter  $b$  introduces a horizontal shift, acting as a hallucination threshold for penalty activation. For this pilot use of the G-PS, parameters  $a$  and  $b$  were set to  $a = 1$  and  $b = 0$  for the most fundamental assessment of the Hallucination Penalty function (ie, no penalty scaling and no threshold). Finally, parameters  $c$  and  $d$  are calibrating factors for the sigmoid function and are set to  $c = 2$  and  $d = 1$  to create bound G-PS values between  $[-1, 1]$ .

$$\text{G-PS} = \text{base score} - \text{hallucination penalty}, \quad 1$$

$$\begin{aligned} \text{Base score} = & (x \times \text{preferred Tx listed}) \\ & + (y \times \text{other preferred Tx listed}), \quad 2 \end{aligned}$$

$$\text{Hallucination penalty} = c \times \frac{e^{a(h-b)}}{1 + e^{a(h-b)}} - d. \quad 3$$

The AI-generated texts were recorded and evaluated for treatment recommendations, including the suggestion to explore a clinical trial, inclusion of academic citations or reference to a pivotal clinical trial, and output length. For this pilot exploratory study, Student's  $t$  test was used to compare outcomes in treatment reporting accuracy, hallucinations, and the G-PS between GPT-3.5 and GPT-4. Statistical analyses were performed using Excel version 16.79. Statistical significance was set to  $P < .05$ . This cross-sectional comparative study was deemed exempt from review and informed consent in accordance with institutional institutional review board policy.

## RESULTS

A total of 160 ChatGPT responses were analyzed. Each output was generated as an NGS report facsimile with at least one treatment recommendation included. GPT-4 generated lengthier responses, with the word count median of 106 for GPT-3.5 (range, 44-232) and 380 for GPT-4 (range, 269-512). GPT-4 had a lower median of four unique treatment options (range, 2-11), and GPT-3.5 had a lower median of five unique treatment options (range, 3-7) with similar rates of recommending participation in a clinical trial (GPT-3.5:

54% and GPT-4: 60%). GPT-4 had 15 runs including a total of 29 citations/references (range, 1-3). Most (52%) were academic citations. None of the included citations or clinical trial references were hallucinated. Examples of prompts and outputs are shown in the Data Supplement (Table S1).

## Treatment Reporting Accuracy

At least one NCCN 1L preferred treatment option was listed in 69% of GPT-3.5 responses versus 100% of GPT-4 responses. GPT-4 demonstrated a significant improvement in preferred treatment reporting (98% v 62%;  $P < .01$ ). Reporting of other recommended treatments was similar between models (GPT-3.5: 48%, GPT-4: 59%;  $P = .13$ ). When the base score weighting was applied, GPT-4 resulted in a significantly higher average base score of 0.90 compared with 0.60 for GPT-3.5 ( $P < .01$ ; Fig 2).

## Hallucinations

On average, GPT-3.5 reported 2.7 hallucinations per response (range, 0-6) versus GPT-4 reporting 1.7 (range, 0-9). In total, GPT-3.5 generated five (6.25%) reports with no hallucinations and GPT-4 generated 19 (23.75%) hallucination-free reports. Overall, GPT-4 exhibited a significantly lower hallucination rate at 34% versus 53% observed in GPT-3.5 ( $P < .01$ ; Fig 3). The hallucination rates translated to lower average hallucination penalties for GPT-4 versus 3.5 (0.56 v 0.75;  $P < .01$ ).

## G-PS

The G-PS, which assesses the overall effectiveness of the models, shifted from a negative score of  $-0.15$  for GPT-3.5 to a positive score of 0.34 for GPT-4 ( $P < .01$ ; Fig 4). GPT-4 performed significantly better than GPT-3.5 for six of the eight mutations, specifically EGFR exon 21 L858R mutation, EGFR exon 19 deletion, ALK rearrangement, BRAF V600E mutation, METex14 skipping mutation, and RET rearrangement. In each of these cases except for ALK rearrangement, GPT-4 resulted in a positive G-PS, whereas GPT-3.5 resulted in a negative G-PS.

## DISCUSSION

We assessed and compared the performance of GPT-3.5 and GPT-4 in generating NGS-like reports with treatment recommendations for 1L aNSCLC with driver mutations. In our assessment, we developed and piloted the G-PS, a novel metric that factors accuracy, relevancy, and hallucinations into the LLM performance assessment. These findings contribute to the literature on LLM chatbots in health care and provide a benchmark for ChatGPT performance in treatment recommendation generation for precision oncology. Notably, our study achieved these results with zero-shot prompting and no task-specific training. Furthermore, it underscores the rapid pace of improvement in LLM models as GPT-4 either outperformed or performed as well as its

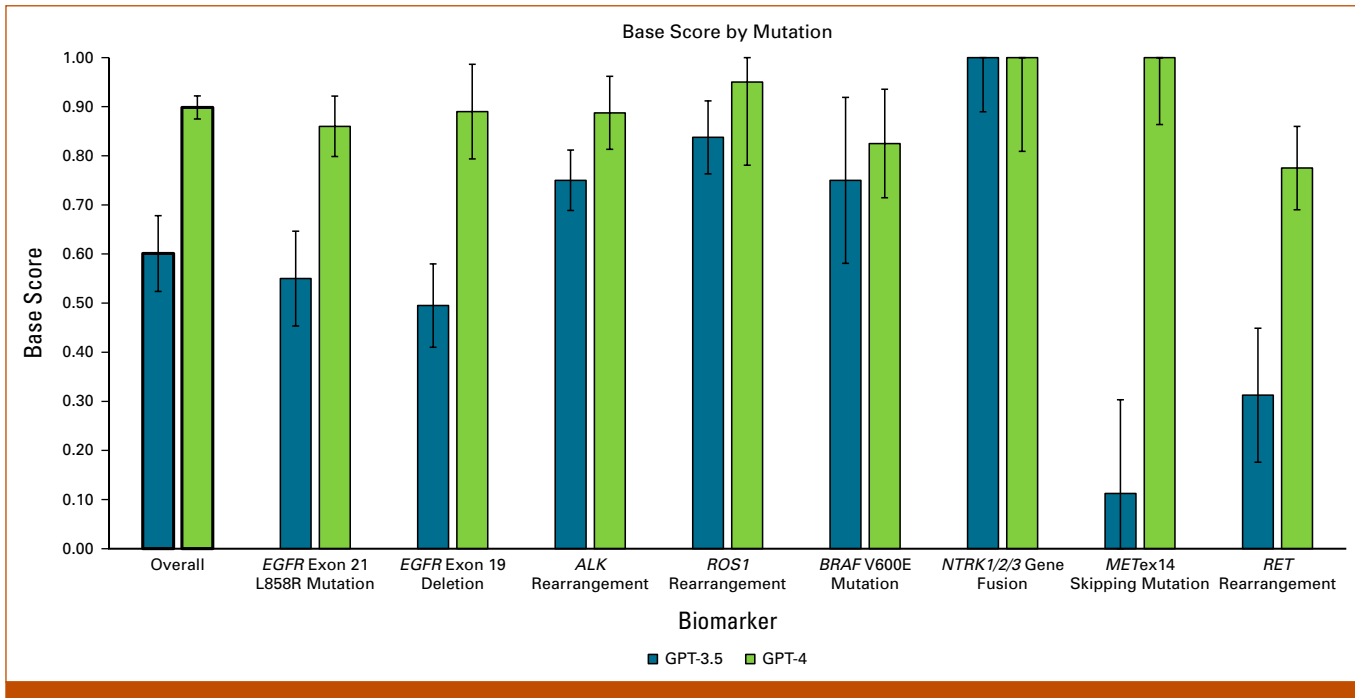


FIG 2. Base accuracy score stratified by the GPT model for overall and individual mutations. GPT, generative pre-trained transformer.

predecessor GPT-3.5 across each driver oncogene in its accuracy, hallucination rate, and G-PS score.

At the time of analysis, both GPT-4 and 3.5 training data were cut off in September 2021.<sup>33</sup> In addition to increased parameters affecting predictive capabilities, GPT-4 has

shown significant advances in understanding and generating contextually appropriate responses, as evidenced by substantial improvements versus GPT-3.5 in publicly available simulated examinations.<sup>34</sup> GPT-4’s improved contextualization was thus crucial to sorting through the vast amount of rapidly developing precision oncology literature

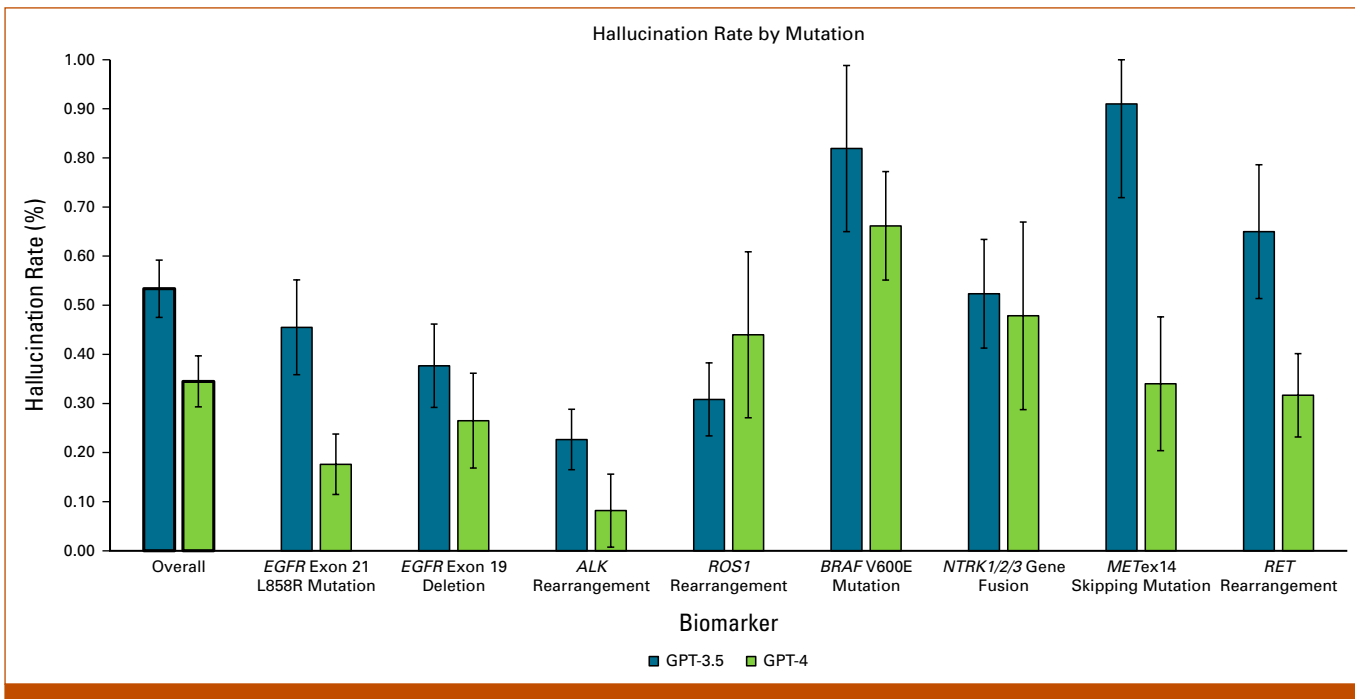
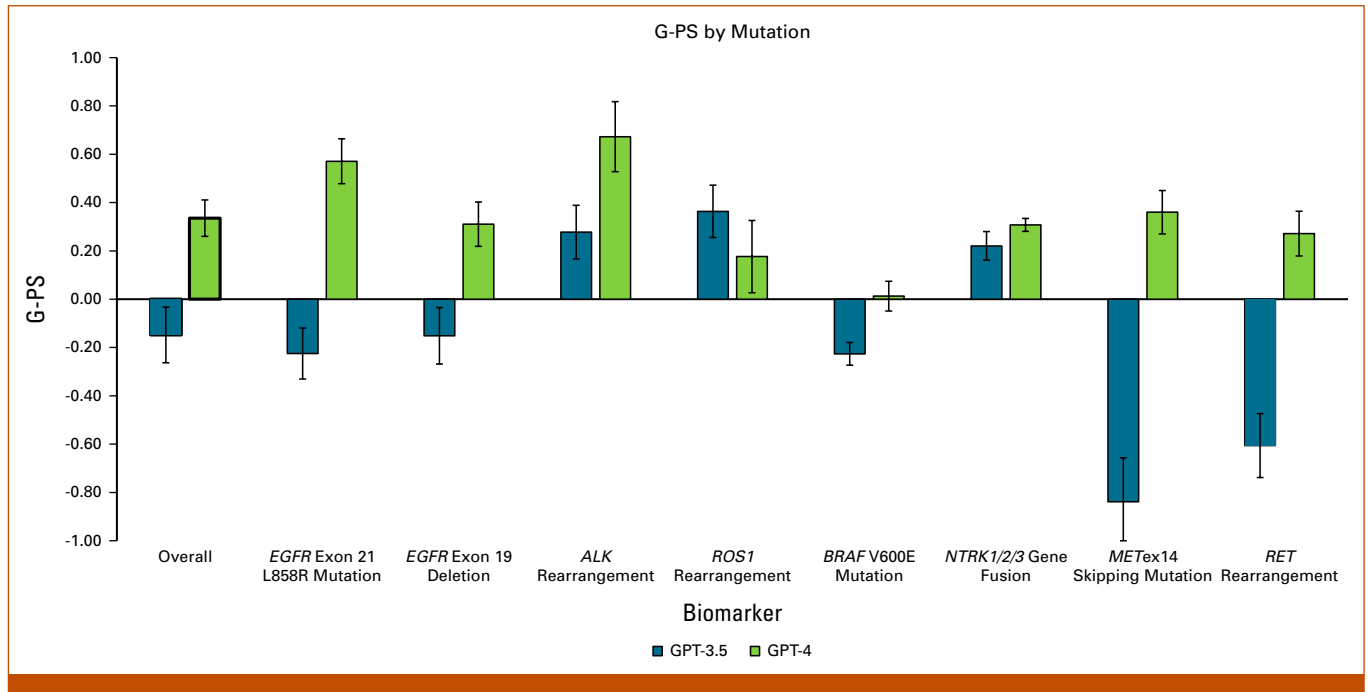


FIG 3. Hallucination rate stratified by the GPT model for overall and individual mutations. GPT, generative pre-trained transformer.





**FIG 4.** G-PS stratified by the GPT model for overall and individual mutations. AI, artificial intelligence; G-PS, Generative AI Performance Score; GPT, generative pre-trained transformer.

available to determine appropriate treatment recommendations in this study. Future research will be needed to validate whether newer models continue to improve their performance relative to national guidelines.

GPT-4 had an improved G-PS over GPT-3.5 in six mutations (EGFR exon 21 L858R mutation, EGFR exon 19 deletion, ALK rearrangement, BRAF V600E mutation, METex14 skipping mutation, and RET rearrangement). Classical EGFR mutations (exon 21 L858R and exon 19 deletion) and ALK rearrangements were the first two biomarkers with FDA approved therapies in aNSCLC (gefitinib in 2003 and crizotinib in 2011). The breadth of EGFR and ALK-related literature and treatment options likely contributed to lower base scores in GPT-3.5, as evidenced by osimertinib being listed as a treatment option in only 6 of 10 outputs for EGFR exon 21 L858R prompts. However, GPT-4's improved contextualization resulted in substantial base score improvements and the lowest GPT-4 hallucination rates for EGFR and ALK alterations among all mutations in this study, with 100% of EGFR exon 21 L858R prompts listing osimertinib as a treatment option.

By contrast, METex14 and RET rearrangements were among the most recent FDA approvals in our study, with capmatinib (May 2020) and tepotinib (February 2021) for METex14 and pralsetinib (Sept 2020) and selpercatinib (May 2020) for RET. Both biomarkers had low base scores and high hallucination rates in the GPT-3.5 model with substantial improvements in both domains with GPT-4. No significant differences in base score, hallucinations, or G-PS were seen with NTRK or ROS1, which received FDA approvals for therapies between

2016 and 2018. These results suggest that GPT-4 performs better relative to GPT-3.5 across both well-established and newly approved biomarkers with improved contextualization, preferred recommendations, and a marked decrease in hallucination rates.

Regarding BRAF, higher hallucination rates in both models may be due to recommending treatment options for BRAF-mutated melanoma, such as cobimetinib, single-agent immunotherapy, or dual immunotherapy. We hypothesize that as a biomarker which first received FDA approvals in melanoma, BRAF's breadth of literature across tumor types likely contributed to the high rates of hallucination with regard to NSCLC treatment. These findings highlight the susceptibility of LLMs to hallucinations especially in a field such as precision medicine, which is becoming more tumor-agnostic and molecularly driven.

G-PS introduces a combined score which factors in accuracy, relevancy, and hallucinations. While most literature on assessment of LLM performance in the medical field has focused on accuracy,<sup>15-18,20</sup> our study is one of the first to factor in this relevancy. In this study, the G-PS base score had a weight of 0.75 for preferred versus 0.25 for other treatment recommendations which were assigned on the basis of the authors' assessment of relative value of preferred treatment recommendations in aNSCLC. However, these values can be modified for future studies on the basis of relative importance of preferred versus other treatment recommendations in other fields. In our analysis, no changes in significance by mutation cross-model analysis were noted with a more aggressive penalty, even weighing of preferred and other

preferred, or both (Data Supplement, Figs S1–S3). In addition, our study is among the first to assess the hallucinations in LLMs relative to decision making. In diseases such as aNSCLC, where treatment selection and timing are crucial to outcomes, there is little if any room for incorrect recommendations. While our study showed a reduced hallucination rate with GPT-4 compared with GPT-3.5 (34% v 53%), the persistence of hallucinations remains a limitation to their clinical implementation.

Additional studies are needed to validate the G-PS metric across other domains including precision oncology and nononcology specialties. Scaled from –1 (all hallucinations) to 1 (all correct treatments), G-PS provides a numerical assessment of LLM performance relative to national guidelines. In addition to the significant improvement in G-PS between GPT-4 and 3.5, the score also changed to positive from negative, indicating a trend toward more correct recommendations and fewer hallucinations. Furthermore, GPT-4 included citations or clinical trial references in 19% of responses, all of which were found to be accurate and relevant. As models improve, such source transparency can help clinicians appraise LLM-provided information. Studies are needed to assess clinician perceptions of LLM performance with respect to accuracy, relevancy, and hallucinations and to determine an acceptable threshold of a model's G-PS before clinical use.

This study has several strengths and limitations. Strengths include a robust zero-shot methodology with multiple generated runs for each mutation. Furthermore, we devised a novel and user-friendly scoring methodology that integrated the evaluation of hallucinations into the overall performance of the GPT model on the defined task. By using a publicly accessible technology in the form of ChatGPT, this work outlines an approachable methodology to assessing LLM performance. The primary limitation of this small sample pilot of the G-PS is the narrow focus on ChatGPT 1L driver oncogene-mutated aNSCLC treatment recommendations, which may limit generalizability to other areas of oncology and other disease types. Further investigation using other LLMs or with modifications of parameters, such as temperature (ie, the degree of output randomness), is warranted.<sup>35</sup> However, the G-PS was designed such that it is approachable, adaptable, and broadly applicable to a

multitude of generative AI and GPT performance assessments as the parameters for base score (ie, accuracy) and hallucination penalty are defined by the user on the basis of relevant context. Secondly, we defined 'treatment recommendations' as any interventions included in the output phrased in a way that could reasonably be interpreted as an option for treatment. This approach tended to be more punitive to the GPT models and might have overestimated hallucinations (Data Supplement, Fig S1). However, in the context of assessing generative AI in health care, low thresholds for hallucinations are essential to ensure patient safety. Third, given that G-PS is a weighted sum of treatment options, the score has the potential to favor tumor types with a larger number of correct treatment options versus types with more limited options. Future larger studies using G-PS should thus evaluate the performance of the scoring system across diseases with a larger range of treatment options to assess for any biases.

Our study underscores the potential of generative AI, like GPT-4, to become a part of precision oncology care, especially for aNSCLC with driver mutations. By introducing the G-PS, we offer a novel approach for evaluating LLMs, emphasizing the importance of detecting hallucinations in performance metrics. In a field where treatment decisions have critical implications, there should be little to no room for hallucinations in clinical decision-making support tools. The application of generative AI tools has the potential to enhance health care provider decision making and optimize patient care; however, further oncology-specific training and validation of models are needed before their use in clinical practice. Future areas of research include assessing the performance of ChatGPT and other LLMs with G-PS in health care fields outside of aNSCLC including nononcology disease and the impact of prompt engineering on the sensitivity of the G-PS. Importantly, while our study assessed the difference between GPT-4 and 3.5, we did not assess whether either models' performance was acceptable to oncologists for use in clinical decision making. We call for more research on physicians' perceptions of LLMs, including thresholds for accuracy, relevancy, and hallucinations. Our findings contribute to the field of AI in health care, highlighting the necessity for validation and performance refinement of LLMs, especially in rapidly evolving fields such as precision oncology.

## AFFILIATIONS

<sup>1</sup>University of Illinois Chicago, Chicago, IL

<sup>2</sup>University of Washington, Seattle, WA

<sup>3</sup>Kirkland & Ellis, Chicago, IL

## CORRESPONDING AUTHOR

Zacharie Hamilton, BA; e-mail: zhamil3@uic.edu.

## PRIOR PRESENTATION

Presented in part at the American Society of Clinical Oncology Annual Meeting, Chicago, IL, June 2-6, 2023.

## SUPPORT

Supported by the UIC Institute for Equitable Health Data Science Research. R.H.N. is a recipient of the Robert A. Winn Diversity in Clinical Trials Career Development Award, funded by Bristol Myers Squibb Foundation.

## AUTHOR CONTRIBUTIONS

**Conception and design:** Zacharie Hamilton, Zhengjia Chen, Noor Naffakh, Vijayakrishna K. Gadi, Christopher Bun, Ryan H. Nguyen  
**Financial support:** Vijayakrishna K. Gadi, Ryan H. Nguyen  
**Administrative support:** Vijayakrishna K. Gadi  
**Collection and assembly of data:** Zacharie Hamilton, Ryan H. Nguyen  
**Data analysis and interpretation:** All authors  
**Manuscript writing:** All authors  
**Final approval of manuscript:** All authors  
**Accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to [www.asco.org/rwc](http://www.asco.org/rwc) or [ascopubs.org/cci/author-center](http://ascopubs.org/cci/author-center).

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](#)).

**Zacharie Hamilton**  
**Consulting or Advisory Role:** 3rdEyeBio

**Noor Naffakh**  
**Honoraria:** Pharmacy Times

**Natalie M. Reizine**  
**Consulting or Advisory Role:** Tempus, Curium Pharma  
**Speakers' Bureau:** EMD Serono, AstraZeneca, Merck, Tempus, Janssen Oncology

**Frank Weinberg**  
**Consulting or Advisory Role:** Regeneron, Jazz Pharmaceuticals, Tempus  
**Speakers' Bureau:** Amgen, Regeneron, Tempus, AstraZeneca

**Shikha Jain**  
**Stock and Other Ownership Interests:** Doximity  
**Honoraria:** Tempus, Healio, Magellan Health, Novartis, Genentech  
**Consulting or Advisory Role:** Healio, Magellan Health  
**Speakers' Bureau:** Practicing Clinician Exchange  
**Travel, Accommodations, Expenses:** MJH Life Sciences

**Vijayakrishna K. Gadi**  
**Stock and Other Ownership Interests:** Sengine Precision Medicine, Novilla, 3rdEyeBio, Phoenix Molecular Designs, New Equilibrium Biosciences, Tahoma Therapeutics, Emerging Markets Cancer Ignition Fund  
**Consulting or Advisory Role:** Puma Biotechnology, Hologic, Seagen/Pfizer, Stemline Therapeutics, Gilead Sciences, AstraZeneca  
**Speakers' Bureau:** Seagen, Hologic, Puma Biotechnology, Stemline Therapeutics  
**Research Funding:** Agendia (Inst), Tizona Therapeutics, Inc, Illumina  
**Travel, Accommodations, Expenses:** Seagen, Genentech/Roche, Puma Biotechnology  
**Open Payments Link:** <https://openpaymentsdata.cms.gov/physician/2511>

**Christopher Bun**  
**Employment:** CancerIQ  
**Stock and Other Ownership Interests:** Wild Type Advocates

**Ryan H. Nguyen**  
**Consulting or Advisory Role:** Merck, Novartis, Regeneron  
**Speakers' Bureau:** Merck  
**Research Funding:** Exelixis  
**Travel, Accommodations, Expenses:** Merck

No other potential conflicts of interest were reported.

## REFERENCES

- Nagl L, Pall G, Wolf D, et al: Molecular profiling in lung cancer. *Memo* 15:201-205, 2022
- National Comprehensive Cancer Network: Non-small cell lung cancer (version 3.2023). 2023. <https://www.nccn.org>
- Morton C, Sarker D, Ross P: Next-generation sequencing and molecular therapy. *Clin Med* 23:65-69, 2023
- Burns L, Jani C, Radwan A, et al: Implementation challenges and disparities in molecular testing for patients with stage IV NSCLC: Perspectives from an urban safety-net hospital. *Clin Lung Cancer* 24:e69-e77, 2023
- Molina-Vila MA, Mayo-de-las-Casas C, Garzón-Ibáñez M, et al: Annotating the next generation sequencing report. *Precis Cancer Med* 3:6, 2020
- Gray SW, Park ER, Najita J, et al: Oncologists' and cancer patients' views on whole-exome sequencing and incidental findings: Results from the CanSeq study. *Genet Med* 18:1011-1019, 2016
- Howlader N, Forjaz G, Mooradian MJ, et al: The effect of advances in lung-cancer treatment on population mortality. *N Engl J Med* 383:640-649, 2020
- Sadik H, Pritchard D, Keeling DM, et al: Impact of clinical practice gaps on the implementation of personalized medicine in advanced non-small-cell lung cancer. *JCO Precis Oncol* 10.1200/PO.22.00246
- Schmid S, Jochum W, Padberg B, et al: How to read a next-generation sequencing report—What oncologists need to know. *ESMO Open* 7:100570, 2022
- Murciano-Goroff YR, Suehnholz SP, Drilon A, et al: Precision oncology: 2023 in review. *Cancer Discov* 13:2525-2531, 2023
- Liu Y, Han T, Ma S, et al: Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiol* 1:100017, 2023
- Egli A: ChatGPT, GPT-4, and other large language models: The next revolution for clinical microbiology? *Clin Infect Dis* 77:1322-1328, 2023
- Rao A, Pang M, Kim J, et al: Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv* 10.1101/2023.02.21.23285886
- Liu S, Wright AP, Patterson BL, et al: Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc* 30:1237-1245, 2023
- Kung TH, Cheatham M, Medenilla A, et al: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2:e0000198, 2023
- Ali R, Tang OY, Connolly ID, et al: Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* 93:1353-1365, 2023
- Johnson D, Goodman R, Patrinely J, et al: Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the chat-GPT model. *Res Sq* 10.21203/rs.3.rs-2566942/v1
- Eriksen AV, Möller S, Ryg J: Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 1:Alp2300031, 2023
- Sushil M, Zack T, Mandair D, et al: A comparative study of zero-shot inference with large language models and supervised modeling in breast cancer pathology classification. *arXiv* 10.48550/arXiv.2401.13887
- Pan A, Musheyev D, Bockelman D, et al: Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncol* 9:1437-1440, 2023
- Ji Z, Lee N, Frieske R, et al: Survey of hallucination in natural language generation. *ACM Comput Surv* 55:1-38, 2023
- Brown TB, Mann B, Ryder N, et al: Language models are few-shot learners. *Adv Neural Inf Process Syst* 33:1877-1901, 2020
- Chen S, Kann BH, Foote MB, et al: Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol* 9:1459-1462, 2023
- Singhal K, Azizi S, Tu T, et al: Large language models encode clinical knowledge. *arXiv* 10.48550/arXiv.2212.13138
- Chan C, You K, Chung S, et al: Assessing the usability of GutGPT: A simulation study of an AI clinical decision support system for gastrointestinal bleeding risk. *arXiv* 10.48550/arXiv.2312.10072
- Mihalache A, Popovic MM, Muni RH: Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol* 141:589, 2023
- Strong E, DiGiammarino A, Weng Y, et al: Performance of ChatGPT on free-response, clinical reasoning exams. *MedRxiv* 10.1101/2023.03.24.23287731
- Sushil M, Kennedy VE, Mandair D, et al: CORAL: Expert-curated oncology reports to advance language model inference. *NEJM AI* 1:Aldbp2300110, 2024
- Papineni K, Roukos S, Ward T, et al: Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2002*, pp 311-318



30. Lin CY: ROUGE: A package for automatic evaluation of summaries, in Text Summarization Branches Out. Barcelona, Spain, Association for Computational Linguistics, 2004, pp 74-81. <https://aclanthology.org/W04-1013.pdf>
  31. Bishop PA, Herron RL: Use and misuse of the Likert item responses and other ordinal measures. *Int J Exerc Sci* 8:297-302, 2015
  32. National Comprehensive Cancer Network: Non-small cell lung cancer (version 5.2021). 2021. <https://www.nccn.org>
  33. Rosoł M, Gašior JS, Łaba J, et al: Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination. *Sci Rep* 13:20512, 2023
  34. OpenAI; Achiam J, Adler S, et al: GPT-4 technical report. arXiv [10.48550/arXiv.2303.08774](https://arxiv.org/abs/10.48550/arXiv.2303.08774)
  35. Davis J, Van Bulck L, Durieux BN, et al: The temperature feature of ChatGPT: Modifying creativity for clinical research. *JMIR Hum Factors* 11:e53559, 2024
-