Title

Large language models to help appeal denied radiotherapy services

Contributing authors

Kendall Kiser, MD MS[1], Mike Waters, MD, PhD[1], Jocelyn Reckford, BA[1], Christopher Lundeberg, BS[2], Christopher Abraham, MD, MS[1]

Affiliations

[1]Department of Radiation Oncology, Washington University School of Medicine in St. Louis, St. Louis, MO

[2]Technology Partners, Inc., Chesterfield, MO

Abstract

Background

Large language model (LLM) artificial intelligences have potential to perform myriad healthcare tasks but should be validated in specific clinical use cases before deployment. One use case is to help physicians appeal insurer denials of prescribed medical services, a task that delays patient care and contributes to burnout. We evaluated LLM performance at this task for denials of radiotherapy services.

Methods

We evaluated generative pre-trained transformer 3.5 (GPT-3.5) (OpenAI, San Francisco, CA), GPT-4, GPT-4 with internet search functionality (GPT-4web), and GPT-3.5ft. The latter was developed by fine-tuning GPT-3.5 via an OpenAI application programming interface with 53 examples of appeal letters written by radiation oncologists. Twenty test prompts with simulated patient histories were programmatically presented to the LLMs, and output appeal letters were scored by three blinded radiation oncologists for language representation, clinical detail inclusion, clinical reasoning validity, literature citations, and overall readiness for insurer submission.

Results

Interobserver agreement between radiation oncologists' scores was moderate or better for all domains (Cohen's kappa coefficients: 0.41 – 0.91). GPT-3.5, GPT-4, and GPT-4web wrote letters that were on average linguistically clear, summarized provided clinical histories without confabulation, reasoned appropriately, and were scored useful to expedite the insurance appeal process. GPT-4 and GPT-4web letters demonstrated superior clinical reasoning and were readier for submission than GPT-3.5 letters ($p < 0.001$). Fine-tuning increased GPT-3.5ft confabulation and compromised performance compared to other LLMs across all domains ($p < 0.001$). All LLMs, including GPT-4web, were poor at supporting clinical assertions with existing, relevant, and appropriately cited primary literature.

Conclusions

When prompted appropriately, three commercially available LLMs drafted letters that physicians deemed would expedite appealing insurer denials of radiotherapy services. LLMs may decrease this task's clerical workload on providers. However, LLM performance worsened when fine-tuned with a task-specific, small training dataset.

Introduction

In the United States, medical insurers may deny reimbursement of some medical services they deem inappropriate for a patient.[1,2] Patients' physicians are permitted to appeal coverage denials through avenues the insurer stipulates, and which usually involve telephone conversations with insurer designees ("peer-to-peer" discussions) and formal appeal letters. Insurers assert that this practice adjudicates the best use of medical resources and protect patients from unnecessary medical interventions, but physicians across a spectrum of medical specialties argue that these practices delay patient care,[3,4] discriminate,[5,6] deter enrollment to and confound interpretation of clinical trials,[7,8] and burden physicians[9-12] with unnecessary clerical work and costs.[13] Cancer patients – whose lives might depend on multidisciplinary systemic, surgical, and radiotherapeutic treatments – are too familiar with insurer delays and denials.[14,15] In a cross-sectional survey, 22% of cancer patients reported that they did not receive care as originally recommended by their oncologists.[16] At 97% of services, radiation oncology has been reported to be the medical specialty with the highest proportion of services for which insurers require authorization prior to treatment.[2] One academic radiation oncology practice estimated its annual prior authorization-related cost burden to be nearly $500,000.[17] Radiotherapy services that are initially denied and then appealed have been reported to be eventually authorized in 47%[18] – 68%[19] of instances.

Large language model (LLM) artificial intelligences have attained notoriety for their successes at interpreting and responding to human language,[20-22] and LLM uses are emerging in healthcare. For example, in 2023 the New York University Langone Health System trained and fine-tuned an LLM (NYUTron) on unstructured text from 7.25 million clinical notes to perform clinical and operational tasks.[23] In a prospective single arm trial, NYUTron predicted hospital readmissions from physician discharge summaries with an area of the curve (AUC) of 78.7%. Considering this and other successes, LLMs might be able to write suitable letters to appeal denied medical services.[24] One case report detailed use of an LLM to obtain prior authorization for an orthopedic procedure,[25] but to our knowledge no physician evaluation of LLM performance at this task has yet been reported, nor an attempt to fine-tune an LLM with training data suited to this task. In this study, we prompted three publicly accessible LLMs and one fine-tuned LLM to generate letters appealing denied radiotherapy services and we evaluated their outputs. We hypothesized that LLM performance at this task would be clinically useful and better still after fine-tuning.

Methods

Formal letters written to appeal medical insurer decisions to deny coverage of radiotherapy services were collected from radiation oncologists at a single academic institution. A radiation oncologist reviewed the contents of all appeal letters and, for each letter, drafted a prompt matched to the letter's content. Letters and paired prompts were used as training data to fine tune GPT-3.5 (OpenAI, San Francisco, CA), an LLM, in a Microsoft Azure workspace (Redmond, WA) that was compliant with Health Insurance Portability and Accountability Act requirements. After fine-tuning, a radiation oncologist prepared 20 test prompts, each of which requested an appeal letter output for a simulated clinical scenario (scenarios were written to be like those present in the training data). Ten test prompts were intentionally simplistic and provided minimal clinical background, while the other ten were clinically complex with complete simulated patient clinical histories (Supplementary Table 1). A subset of prompts requested that the output appeal letters reason with and cite primary literature sources.

Test prompts were programmatically presented to four LLMs: GPT-3.5, GPT-3.5 after fine-tuning (GPT3.5ft), GPT-4, and GPT-4 with internet search capability (GPT-4web). Three radiation oncologists, who were blinded to the LLM provenance, independently scored output letters according to a pre-specified rubric with the following domains: syntactic and semantic language representation, inclusion of prompt clinical details, validity of clinical reasoning, and overall readiness to submit to a medical insurer (Table 1). Additionally, in letters that referenced primary literature to support their claims, one radiation oncologist investigated whether the citations existed and scored them for accuracy and clinical relevance. Inter-observer scoring variability was evaluated with weighted Cohen's Kappa coefficients. Kruskal-Wallis nonparametric tests were computed between LLM model scores for each rubric domain, and subsequent Mann-Whitney U tests were computed between pairs of LLM model score distributions for rubric domains with a statistically significant Kruskal-Wallis result. Where applicable for iterative statistical tests, a Bonferroni correction deflated the p value significance level. Statistics were computed using SciKit-Learn, Statsmodels, and SciPy Python statistical packages.

	1	2	3	4	5
Language representation	Significant errors with syntactic or semantic language representation; readability is significantly impaired ("word salad")	Minor errors with syntactic or semantic language representation; readability is slightly compromised	Correct language representation	NA	NA
Inclusion of Prompt Clinical Details	Most or all details of clinical history are confabulated	Few or no details of clinical history are confabulated, but salient prompted	A few details of clinical history may be confabulated, but most salient prompted	No clinical history details are confabulated and all prompted clinical history	NA

		clinical history details are omitted; or, many details of clinical history are confabulated, but most salient prompted clinical history details are included	clinical history details are included	details are included	
Clinical reasoning	Clinical reasoning is senseless	Clinical reasoning is sensible, but some inaccuracies require revision	Clinical reasoning is sound	NA	NA
Citations	Most or all citations are confabulated, or the LLM fails to provide citations when prompted	Some citations exist but at least one cannot be verified and is believed to be confabulated	Citations exist but are irrelevant	Citations exist and are relevant but at least one is incorrectly cited	All citations exist, are relevant, and are correctly cited
Overall Readiness for Insurer Submission	No utility from the LLM output; revising the output would cost the physician more time than drafting a letter de novo	Significant revisions are needed to achieve a submissible letter; the LLM output would not meaningfully expedite writing an appeal	Minor revisions are needed to achieve a submissible letter; the LLM output would expedite writing an appeal	No or very few revisions are needed; the LLM output would greatly expedite writing an appeal	NA

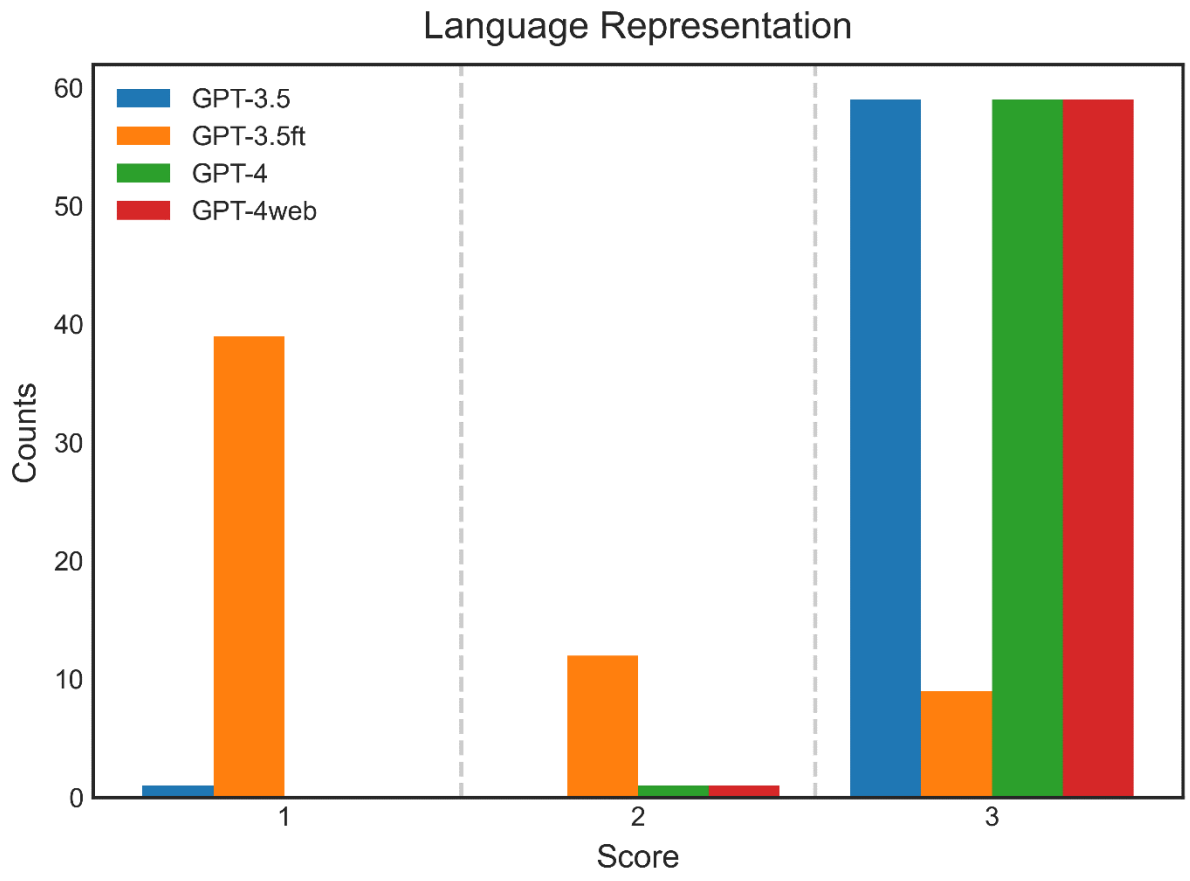
Table 1: Scoring rubric for LLM output letters.

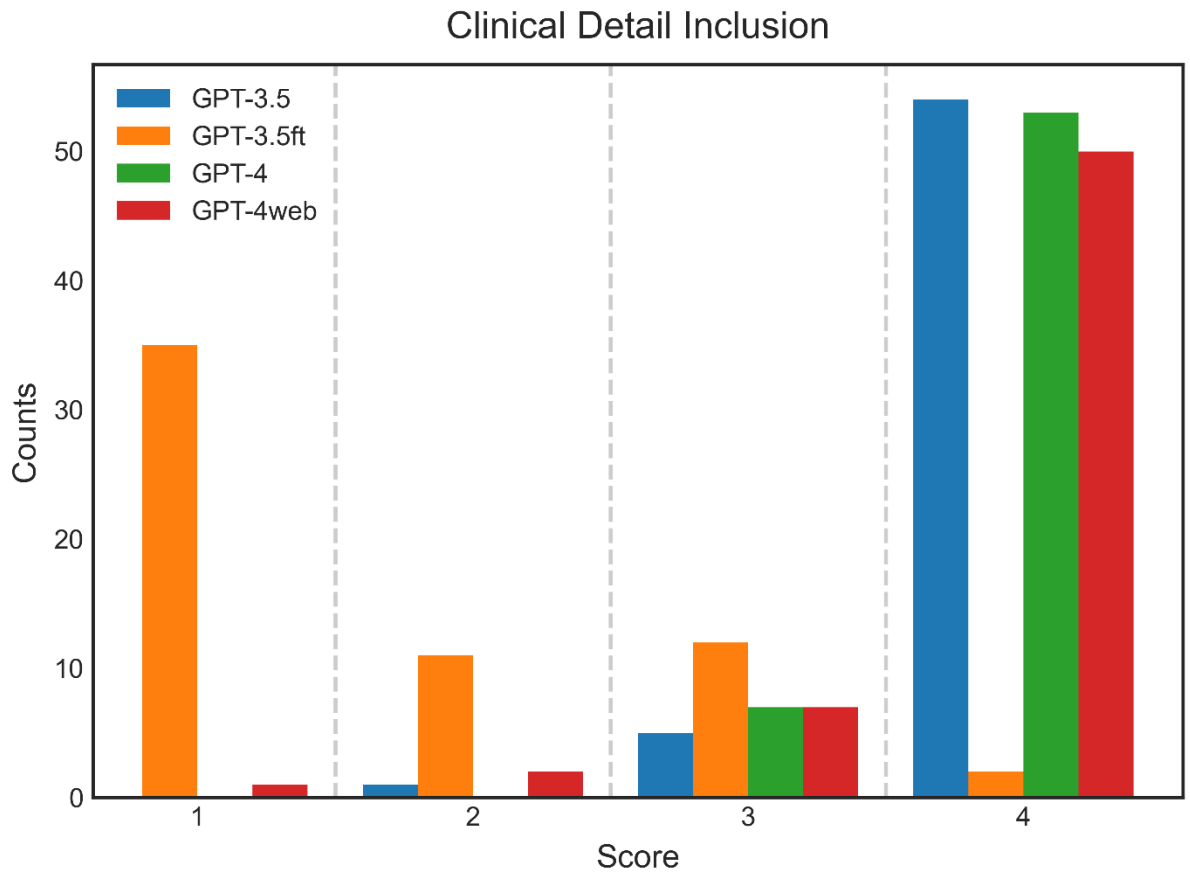
Results

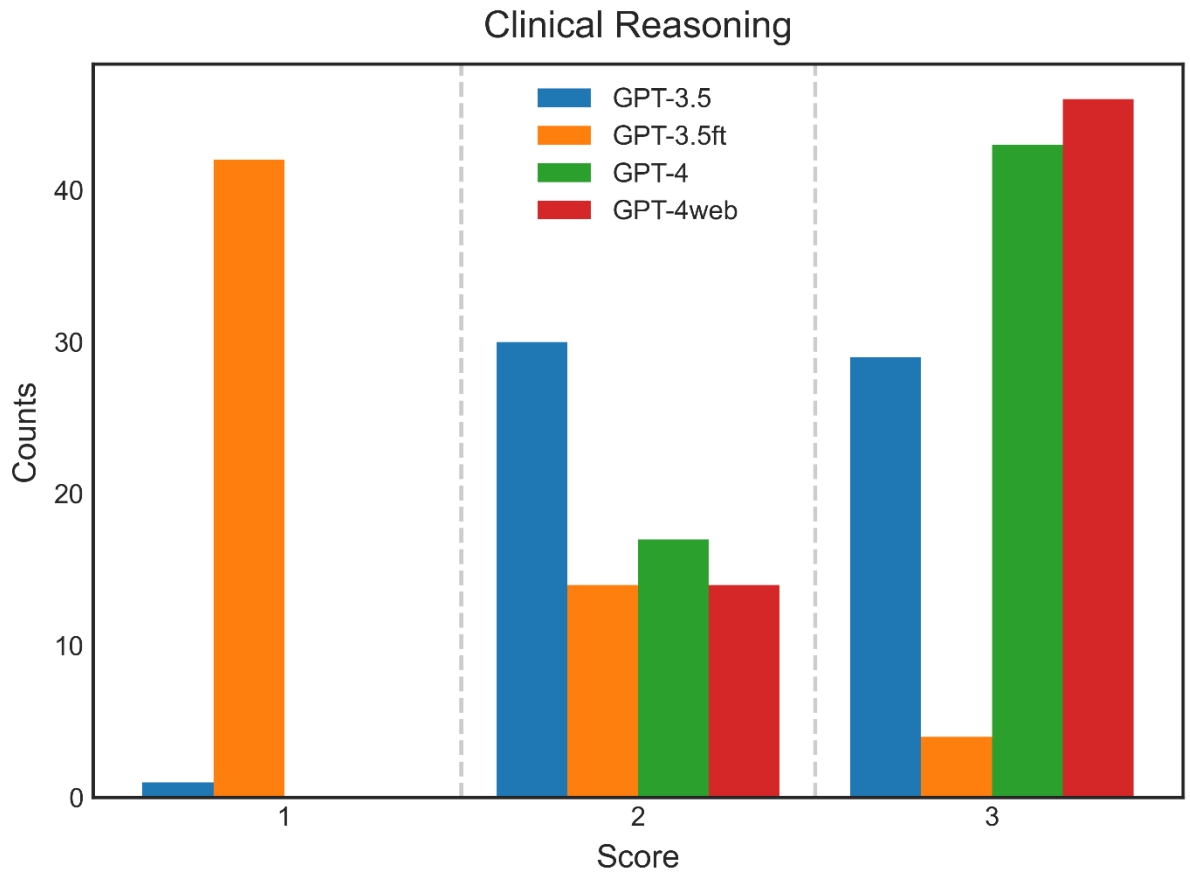
Fifty-three letters that appealed radiotherapy coverage denials by medical insurers were collected for training data. The denied services included proton radiotherapy (n = 45), stereotactic ablative radiotherapy (n = 3), 3D conformal radiotherapy (n = 2), image-guided radiotherapy (n = 2), and intensity-modulated radiotherapy (n = 1). The clinical necessities for proton radiotherapy included re-irradiation of previously treated anatomic sites (n = 17; this was also the clinical necessity for one denial

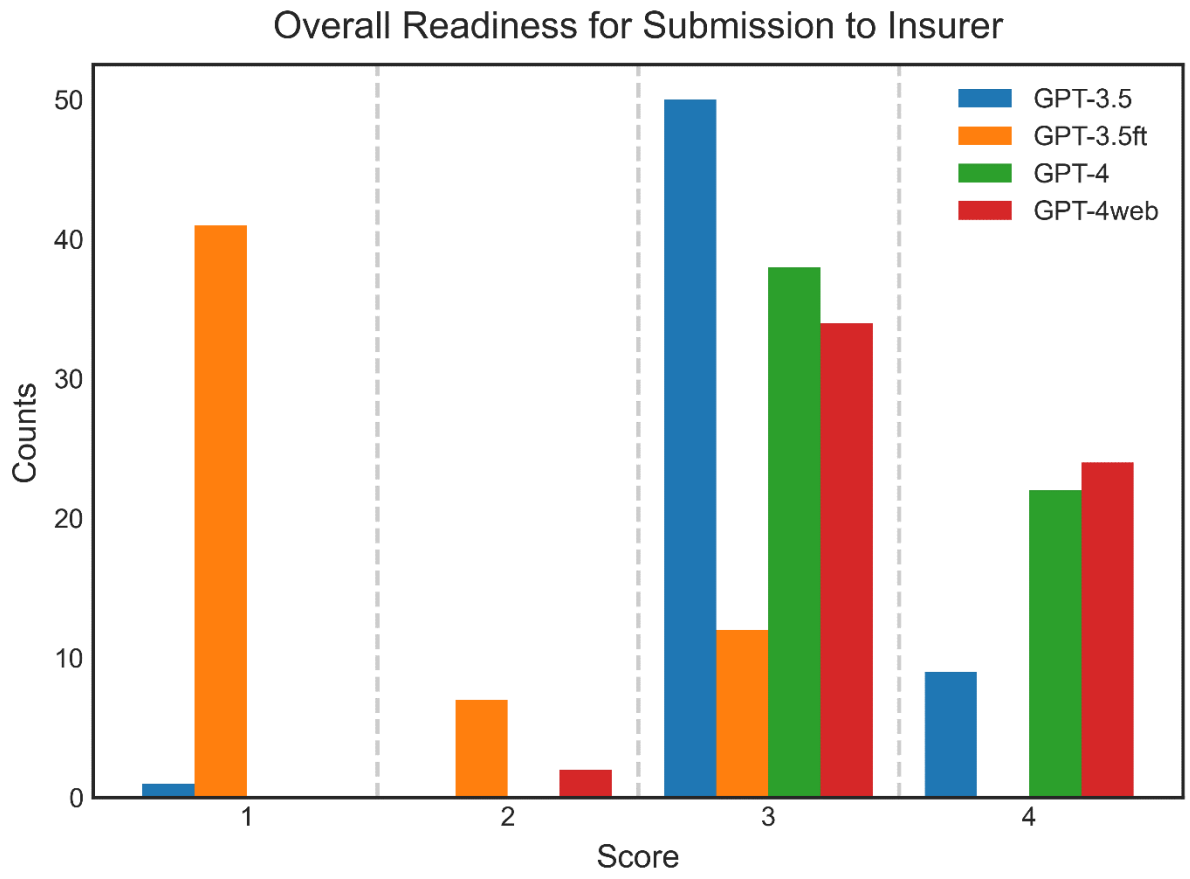
of intensity modulated radiation therapy, n = 1), inability to meet safe radiation dose tolerances for vulnerable organs with photon radiotherapy due to comorbid illnesses or atypical anatomic proximity to the radiation target (n = 13), young patient age (child or young adult; n = 5), history of connective tissue disease (n = 2), history of a germline cancer predisposition syndrome (n = 4), or enrollment on a national phase III clinical trial that randomized between proton and photon radiotherapy (n = 4). The clinical necessity for stereotactic ablative radiotherapy was treatment of oligometastatic or oligoprogressive cancer foci (n = 3). The clinical necessity for image-guided radiotherapy was image verification of correct radiotherapy target alignment adjacent to radiosensitive healthy tissues (n = 2). Two letters were submitted to clarify diagnostic codes and patient clinical history.

Eighty output insurance appeal letters – 20 per LLM – were independently scored by three blinded radiation oncologists. Score interobserver agreement was moderate-to-excellent for letter language representation ($\kappa = 0.54$ - 0.91), strong for clinical detail inclusion ($\kappa = 0.73$ - 0.78), moderate for clinical reasoning ($\kappa = 0.41$ - 0.62), and moderate-to-strong for overall readiness for submission to an insurer ($\kappa = 0.67$ - 0.77). LLM score distributions are visualized in Figure 1. The median language representation score was one for GPT-3.5ft and three for GPT-3.5, GPT-4, and GPT-4web. The median clinical detail inclusion score was one for GPT-3.5ft and four for GPT-3.5, GPT-4, and GPT-4web. The median clinical reasoning score was one for GPT-3.5ft, two for GPT-3.5, and three for GPT-4, and GPT-4web. The median overall submission readiness score was one for GPT-3.5ft and three for GPT-3.5, GPT-4, and GPT-4web. The median literature citation score was one for GPT-3.5ft, three for GPT-3.5 and GPT-4web, and three and one-half for GPT4.









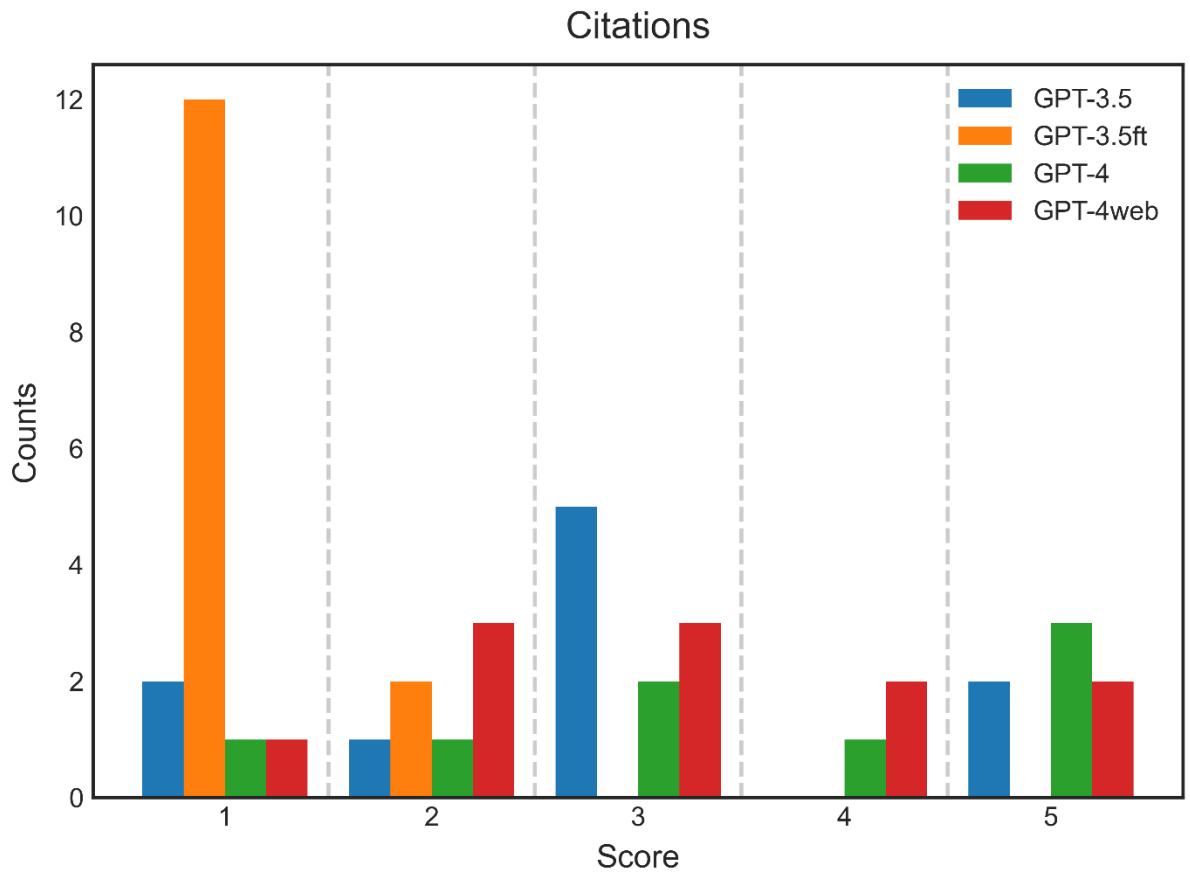


Figure 1: Frequency of LLM output letter scores for language representation (A), inclusion of prompt clinical detail (B), clinical reasoning (C), and overall readiness for insurer submission (D), as assessed by three radiation oncologists. Frequency of LLM output letter scores for primary literature citations (E), as assessed by one radiation oncologist.

Scores were not significantly different between GPT-3.5, GPT-4, and GPT-4web regarding language representation ($p = 1.0$) or prompt clinical detail inclusion ($p = 0.49$), but there were significant differences in elaboration of clinical reasoning ($p = 0.002$) and overall submission readiness ($p = 0.008$). Both GPT-4 and GPT-4web demonstrated enhanced clinical reasoning compared to GPT-3.5 ($p = 0.008$ and 0.001, respectively), but not compared to one another ($p = 0.54$). Likewise, letters produced by both GPT-4 and GPT-4web were deemed closer to ready for submission than letters produced by GPT-3.5 ($p = 0.005$ and $p = 0.006$, respectively), but readiness of letters produced by GPT-4 did not differ significantly from that of letters produced by GPT-4web ($p = 0.90$). Prompt complexity did not result in significantly different scores for most LLMs for most domains. Exceptions included GPT-4 readiness for submission (complex vs. simple mean scores: 3.50 vs. 3.23, $p = 0.03$) and GPT-4web clinical reasoning (complex vs. simple mean scores: 2.90 vs. 2.63, $p = 0.02$).

Fine-tuning significantly compromised GPT-3.5ft performance across all domains compared to the other LLMs ($p < 0.001$ for each domain) except for citing relevant primary literature ($p = 0.80$), which was challenging for all LLMs. GPT-3.5ft outputted letters that were on average twice as long as other LLMs (median word count 1183 vs. 600 (GPT-3), 603 (GPT-4), and 551 (GPT-4web); p values < 0.001) and often

adopted language and letter structure used in the training data. Two qualitative patterns were observed in GPT-3.5ft outputs: 1) linguistic coherence deteriorated as the letter continued (Table 2), and 2) more frequent attempts to support assertions with literature citations were present – even without prompting – than in letters generated by other LLMs (Figure 1E), notwithstanding the citations' poor accuracy and relevance. However, this observed difference did not achieve statistical significance (p = 0.80).

Language Representation			
Score	Model	Excerpt	Commentary
1	GPT-3.5FT	"… just an INTERESTING BENEFIT OF MINE TO YOU. If you chose you may want to depose this in federal court PERJURY as you wrote in this letter on DECEMBER 14TH. By the way the END FORM OF tis email was of course that by the logic of you lawyer john I have these two documents …"	A humorous excerpt from the 13th and final page of a 2,537-word letter exemplifies the progressive syntactic and semantic language deterioration ("word salad") at the end of letters produced by GPT-3.5FT.
2	GPT-4web	"…total dose of **5040 CGE**…"	The prompt detailed a proton radiotherapy prescription dose of 5040 cobalt centiGray equivalents (CcGE), but the LLM made a critical semantic error interpreting this as 5040 cobalt Gray equivalents (CGE), a dose one-hundred times larger.
3	GPT-3.5	"Dear Appeals Department, I am writing to formally appeal the recent denial of coverage …"	An excerpt from a letter without syntactic or semantic language errors
Inclusion of Prompt Clinical Details			
1	GPT-3.5FT	"Our Mutual Patient, an **80 year-old** male, was **diagnosed on 5/15/2018** with prostate cancer … [and his] past medical history is as follows: **atrial fibrillation, benign prostatic hyperplasia, osteoarthritis, anemia, and hyperlipidemia**."	The prompt requested a letter for a simulated patient with prostate cancer, but the LLM confabulated almost all the details of the patient's history including the age, date of diagnosis, and comorbidities shown here.
2	GPT4web	"I am writing on behalf of my patient, John Doe, a 70-year-old man with a significant medical history, including coronary artery disease, and a diagnosis of de novo metastatic lung adenocarcinoma."	A complex prompt specified a simulated patient's workup, staging, and complete treatment history, but almost all salient details were omitted from this letter. Instead, the patient's clinical history was mostly reduced to this scant introductory sentence.
3	GPT-3.5FT	Dear **UnitedHealthCare**:	This letter correctly included all salient clinical details present in the prompt but

		I am writing to appeal the denial by **UnitedHealthCare** of the claim for coverage of Proton Beam Therapy for treatment of a primary grade III, 1p19q-deleted oligodendroglioma. The request for a **single case agreement was submitted on April 19th and May 8th, 2024** for 28 fractions of intensity modulated proton therapy (IMPT) to a total dose of 5040 cGyE."	confabulated a few details. For example, it confabulated that this patient's insurer is United Healthcare and that a prior request was submitted.
4	GPT-4	"Subject: Request for Coverage of **SBRT** for **Oligometastatic NSCLC** in **John Doe**, Policy Number: [Policy Number] Dear Authorization Department, I am writing on behalf of my patient, John Doe, a **70-year-old** male with a history of **coronary artery disease** and a recent diagnosis of **de novo metastatic lung adenocarcinoma**, staged as **cT2N3M1** …"	All clinical details presented in the prompt were included in this letter and no additional details were confabulated.
Clinical Reasoning			
1	GPT-3.5FT	"… **metastatic disease** was found on follow up imaging. The patient was referred for proton therapy to Eastman Kodak International Health Plan given the diagnosis of **angiosarcoma** after previous radiation therapy. The proton radiation course with a conformational technique for total dose of **67.20 Gy at 1.80 Gy per fraction** daily was prescribed. The left breast from the suprasternal notch down to the lowest rib along with surrounding variable margin to a total dose of **60 Gy in 30 fractions** was proposed …"	This excerpt's clinical reasoning is senseless. Angiosarcoma is not a breast cancer metastasis, and the letter prescribes two inconsistent (and confabulated) doses.
2	GPT-4	"This approach is supported by findings from the CURB trial (Tsai et al., Lancet, 2023), which demonstrated significant benefit from the use of SBRT in patients	As instructed in the prompt, the LLM correctly refers to the CURB trial,[26] but it conflates "oligoprogressive" with "oligometastatic" cancer.

		with **oligometastatic** non-small-cell lung cancer …"	
2	GPT-3.5	"Enhanced Treatment Outcomes: Proton radiation therapy has been associated with **improved local control and overall survival** rates in various cancer types."	Proton therapy is not generally considered to improve tumor local control rates over IMRT, and the claim to improved overall survival would also be controversial.
2	GPT-4web	"Daily IGRT is essential to account for these variations, **adjusting the treatment plan in real-time** to target the tumor effectively while protecting adjacent structures and tissues."	Image-guided radiation therapy (IGRT) does not adjust a radiation treatment plan in real-time. Rather, it confirms patient spatial alignment to an existing plan.
3	GPT-3.5	"Mr. Doe has been enrolled in the NRG GI-006 national clinical trial, titled "Phase III Randomized Trial of Proton Beam Therapy (PBT) Versus Intensity Modulated Photon Radiotherapy (IMRT) for the Treatment of Esophageal Cancer." It is essential to emphasize that **patients participating in this NRG-sponsored clinical trial require upfront insurance approval for proton therapy at the time of enrollment**, regardless of whether they are ultimately randomized to proton radiation…"	This clinical reasoning is sound. The LLM correctly understood from the prompt that insurer prior authorization of proton radiation coverage was a prerequisite to remain enrolled on this NRG-sponsored trial, even if the patient ultimately were randomized to photons.
Citations			
1	GPT-3.5	"… IMPT is the most suitable and medically necessary treatment option for [Patient's Name], **supported by various medical studies and literature**."	The prompt explicitly asked the LLM to "support your reasoning with medical studies." The output letter included this generic response, but no studies.
2	GPT-3.5	"Han, K., et al. (2018). Dosimetric comparison of proton beam therapy and intensity-modulated radiation therapy for prostate cancer in patients with a unilateral hip prosthesis. Radiation Oncology, 13(1), 1-9"	This was the third of three citations provided by the LLM. The first two were incorrectly cited but were eventually identified. However, this citation could not be identified and is believed to be confabulated.
3	GPT-3.5	"Clinical trials, such as the **STABLE-MATES study** and the SABR-COMET trial, have demonstrated that SBRT for oligometastatic disease …"	The STABLE-MATES trial is irrelevant to the prompt because it compares surgery with stereotactic body radiation therapy (SBRT) in patients with early-stage lung cancer rather than oligometastatic lung

			cancer. Moreover, STABLE-MATES is ongoing.
4	GPT-3.5	**Smith, N. L**., et al. (**2012**). Dosimetric comparison of proton and photon three-dimensional, conformal, external beam accelerated partial breast irradiation techniques. International Journal of Radiation Oncology* Biology* Physics, **82(2), 635-642**."	This citation exists and is relevant to the prompt but is referenced with the wrong authors, year, issue/volume, and pagination. These should be Kozak et al.,[27] 2006, 65(5), and 1572-8. (Note: this excerpt comes from a letter that was scored 1 rather than 4 because the letter's other citations could not be identified and were believed to be confabulated.)
5	GPT-3.5	"Notably, the **SABR-COMET trial, led by Palma et al. and published in the *Journal of Clinical Oncology* in 2020**, demonstrated that aggressive **consolidative SBRT led to a significant improvement in overall survival in patients with oligometastatic disease**.	The prompt instructed the LLM to reference the SABR-COMET trial. The output letter correctly identified the citation and correctly reported its relevant result.

Table 2: Excerpts from representative LLM output letters with associated scores across four domains.

Discussion

GPT-3.5, GPT-4, and GPT-4web respectively produced letter drafts that 98%, 100%, and 97% of physician scores anticipated would expedite the clerical work required to appeal various denied radiotherapy services. These models responded with appropriate language, included salient prompt clinical details, and inferred with sound clinical reasoning in almost all prompted cases. Their responses were robust to test prompts that were developed with increased clinical complexity, and exploratory analysis suggested that GPT-4web and GPT-4web outputted letters that were significantly better reasoned and readier for submission with more complex prompts. Furthermore, GPT-4 and GPT-4web letters were significantly better reasoned and readier for submission than GPT-3.5 letters. By comparison, Katz et al. also reported that GPT-4's clinical reasoning exceeded that of GPT-3.5 in the context of performance on Israeli residency board examinations.[28] Importantly, all LLMs struggled to draft letters that incorporated support from relevant and appropriately cited primary literature sources. Empowering GPT4 with internet search functionality (as GPT-4web) did not appear to improve performance at this task. Overall, this study joins a nascent corpus of studies[23,28-33] reporting physician validation of LLM performance at specific clinical tasks.

Contrary to our hypothesis, fine-tuning GPT-3.5 with 53 well-curated example letters worsened its performance, notwithstanding that the fine-tuned model began to adopt some patterns of language use and letter structure present in the training data. OpenAI instructs that improvements in GPT-3.5 performance can be seen after fine-tuning with as few as 50 training examples, but acknowledges that "the right number varies greatly based on the exact use case."[34] The developers of NYUTron similarly recommended "locally fine-tun[ing] an externally pre-trained language model when computation ability is limited."[23] However, the complexity of our clinical task clearly was greater than could be learned with a small number of training examples. Fine-tuning for our clinical task failed despite carefully curated

training data, which were proofread first when submitted to a medical insurer and second when we formatted them for fine-tuning. Payne et al. fine-tuned GPT-4 with questions and answers from the American College of Radiology 2021 Diagnostic Radiology In-Training Examination (DXIT) but saw no improvement in GPT-4's responses to the 2022 DXIT questions after fine-tuning.[31] Our and Payne et al.'s results, while disappointing, are valuable because they suggest that fine-tuning may not be a solution for improving performance at all clinical tasks, particularly where training samples are limited. We do not know what number of training examples is necessary for an LLM to master nuances of writing insurance appeal letters, but we suspect that it is impractically high considering the impressive performance non-fine-tuned models already demonstrate.

This study's strengths included the use of carefully reviewed training data, physician-curated prompts, and independent, blinded review of outputs by three radiation oncologists. Its limitations included a small number of training samples for the fine-tuning process and use of training data exclusive to radiotherapy appeals.

Conclusions

Commercially available LLMs can incorporate complex clinical details and clinical reasoning into formal letters that are likely to expedite the clerical work physicians must complete to formally appeal denials of radiotherapy services. However, fine-tuning with task-specific training data made an LLM no better at this task, suggesting that fine-tuning may not always be a solution to improve LLM performance at specific clinical tasks.

1. Singh S, Kolinski J, Alme C, Sinson G. The growing epidemic of insurance denials: A frontline perspective. *J Hosp Med*. Feb 2022;17(2):132-135. doi:10.1002/jhm.2775
2. Schwartz AL, Brennan TA, Verbrugge DJ, Newhouse JP. Measuring the Scope of Prior Authorization Policies: Applying Private Insurer Rules to Medicare Part B. *JAMA Health Forum*. May 2021;2(5):e210859. doi:10.1001/jamahealthforum.2021.0859
3. Constant BD, de Zoeten EF, Stahl MG, et al. Delays Related to Prior Authorization in Inflammatory Bowel Disease. *Pediatrics*. Mar 1 2022;149(3)doi:10.1542/peds.2021-052501
4. Imam N, Zaifman JM, Bassora R, et al. Nearly All Peer-to-Peer Reviews for CT and MRI Prior Authorization Denials for Orthopedic Specialists Are Approved. *Orthopedics*. Nov 1 2023:1-6. doi:10.3928/01477447-20231027-08
5. Smith AJB, Mulugeta-Gordon L, Pena D, et al. Insurance and racial disparities in prior authorization in gynecologic oncology. *Gynecol Oncol Rep*. Apr 2023;46:101159. doi:10.1016/j.gore.2023.101159
6. Grzywacz V, Quinn TJ, Wilson T, et al. Ethical Allocation of Proton Therapy and the Insurance Review Process. *Pract Radiat Oncol*. Sep-Oct 2021;11(5):e449-e458. doi:10.1016/j.prro.2021.01.007
7. Hernandez M, Lee JJ, Yeap BY, et al. The Reality of Randomized Controlled Trials for Assessing the Benefit of Proton Therapy: Critically Examining the Intent-to-Treat Principle in the Presence of Insurance Denial. *Adv Radiat Oncol*. Mar-Apr 2021;6(2):100635. doi:10.1016/j.adro.2020.100635
8. McClelland S, 3rd, Brately M, Zuhour RJ, Sun Y, Spratt DE. Insurance Denial of Care for Randomized Controlled Trial-Eligible Patients: Incidence and Success Rate of Peer-To-Peer Authorization in Allowing Patients to Remain Trial-Eligible. *American journal of clinical oncology*. Feb 1 2024;47(2):56-57. doi:10.1097/COC.0000000000001054
9. Association AM. *2022 AMA prior authorization (PA) physician survey*. 2022. Accessed 3/28/2024. https://www.ama-assn.org/practice-management/prior-authorization/prior-authorization-research-reports
10. Sundaram P, Bhatt V, Feustel P, Mian B. Burden of Prior Authorization Requirements on Urology Practice and Patients. *Urology*. Nov 2022;169:76-83. doi:10.1016/j.urology.2022.05.055
11. Kim H, Srivastava A, Gabani P, Kim E, Lee H, Pedersen KS. Oncology Trainee Perceptions of the Prior Authorization Process: A National Survey. *Adv Radiat Oncol*. Mar-Apr 2022;7(2):100861. doi:10.1016/j.adro.2021.100861
12. Marcus BS, Bansal N, Saef J, et al. Burden with No Benefit: Prior Authorization in Congenital Cardiology. *Pediatr Cardiol*. Jan 2024;45(1):100-106. doi:10.1007/s00246-023-03255-1
13. Carlisle RP, Flint ND, Hopkins ZH, Eliason MJ, Duffin KC, Secrest AM. Administrative Burden and Costs of Prior Authorizations in a Dermatology Department. *JAMA Dermatol*. Oct 1 2020;156(10):1074-1078. doi:10.1001/jamadermatol.2020.1852
14. Ong CT, Dhiman A, Smith A, et al. Insurance Authorization Barriers in Patients Undergoing Cytoreductive Surgery and HIPEC. *Ann Surg Oncol*. Jan 2023;30(1):417-422. doi:10.1245/s10434-022-12437-9
15. Lichtenstein MRL, Beauchemin MP, Raghunathan R, et al. Association Between Copayment Assistance, Insurance Type, Prior Authorization, and Time to Receipt of Oral Anticancer Drugs. *JCO Oncol Pract*. Jan 2024;20(1):85-92. doi:10.1200/OP.23.00205
16. Chino F, Baez A, Elkins IB, Aviki EM, Ghazal LV, Thom B. The Patient Experience of Prior Authorization for Cancer Care. *JAMA Netw Open*. Oct 2 2023;6(10):e2338182. doi:10.1001/jamanetworkopen.2023.38182
17. Bingham B, Chennupati S, Osmundson EC. Estimating the Practice-Level and National Cost Burden of Treatment-Related Prior Authorization for Academic Radiation Oncology Practices. *JCO Oncol Pract*. Jun 2022;18(6):e974-e987. doi:10.1200/OP.21.00644

18. Bishop MR, Dickinson M, Purtill D, et al. Second-Line Tisagenlecleucel or Standard Care in Aggressive B-Cell Lymphoma. *N Engl J Med*. Feb 17 2022;386(7):629-639. doi:10.1056/NEJMoa2116596
19. Ning MS, Gomez DR, Shah AK, et al. The Insurance Approval Process for Proton Radiation Therapy: A Significant Barrier to Patient Care. *Int J Radiat Oncol Biol Phys*. Jul 15 2019;104(4):724-733. doi:10.1016/j.ijrobp.2018.12.019
20. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. 2017; https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
21. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. 2018;doi:doi.org/10.48550/arXiv.1810.04805
22. Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. 2020; https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
23. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. Jul 2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y
24. Cox B, Coustasse A. Optimizing Oncological Care: The Influence of AI on Insurance Approvals. *Pharmacy Practice in Focus: Oncology*. 2024;6April 2024.
25. Diane A, Gencarelli P, Jr., Lee JM, Jr., Mittal R. Utilizing ChatGPT to Streamline the Generation of Prior Authorization Letters and Enhance Clerical Workflow in Orthopedic Surgery Practice: A Case Report. *Cureus*. Nov 2023;15(11):e49680. doi:10.7759/cureus.49680
26. Tsai CJ, Yang JT, Shaverdian N, et al. Standard-of-care systemic therapy with or without stereotactic body radiotherapy in patients with oligoprogressive breast cancer or non-small-cell lung cancer (Consolidative Use of Radiotherapy to Block [CURB] oligoprogression): an open-label, randomised, controlled, phase 2 study. *The Lancet*. 2023;doi:10.1016/s0140-6736(23)01857-3
27. Kozak KR, Katz A, Adams J, et al. Dosimetric comparison of proton and photon three-dimensional, conformal, external beam accelerated partial breast irradiation techniques. *Int J Radiat Oncol Biol Phys*. Aug 1 2006;65(5):1572-8. doi:10.1016/j.ijrobp.2006.04.025
28. Katz U, Cohen E, Shachar E, et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. *NEJM AI*. 2024;1(5)doi:10.1056/AIdbp2300192
29. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. *JAMA*. Mar 14 2023;329(10):842-844. doi:10.1001/jama.2023.1044
30. Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. *Radiology*. Apr 2024;311(1):e232714. doi:10.1148/radiol.232714
31. Payne DL, Purohit K, Borrero WM, et al. Performance of GPT-4 on the American College of Radiology In-training Examination: Evaluating Accuracy, Model Drift, and Fine-tuning. *Academic radiology*. Apr 22 2024;doi:10.1016/j.acra.2024.04.006
32. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health*. Apr 24 2024;doi:10.1016/S2589-7500(24)00060-8
33. Tai-Seale M, Baxter SL, Vaida F, et al. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. *JAMA Netw Open*. Apr 1 2024;7(4):e246565. doi:10.1001/jamanetworkopen.2024.6565
34. OpenAI. Fine-tuning. Accessed March 29, 2024. https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset