

# A machine learning model using clinical notes to identify physician fatigue

Received: 9 January 2024

Accepted: 6 June 2025

Published online: 01 July 2025

 Check for updates

Chao-Chun Hsu  <sup>1</sup>, Ziad Obermeyer  <sup>2</sup> & Chenhao Tan  <sup>1</sup> 

Clinical notes should capture important information from a physician-patient encounter, but they may also contain signals indicative of physician fatigue. Using data from 129,228 emergency department (ED) visits, we train a model to identify notes written by physicians who are likely to be tired: those who worked ED shifts on at least 5 of the prior 7 days. In a hold-out set, the model accurately identifies notes written by such high-workload physicians. It also flags notes written in other settings with high fatigue: overnight shifts and high patient volumes. When the model identifies signs of fatigue in a note, physician decision-making for that patient appears worse: yield of testing for heart attack is 19% lower with each standard deviation increase in model-predicted fatigue. A key feature of notes written by fatigued doctors is the predictability of the next word, given the preceding context. Perhaps unsurprisingly, because word prediction is the core of how large language models (LLMs) work, we find that predicted fatigue of LLM-written notes is 74% higher than that of physician-written ones, highlighting the possibility that LLMs may introduce distortions in generated text that are not yet fully understood.

Physicians write notes about patient encounters to convey information and summarize their thinking. Researchers increasingly use these notes to make inferences about patients' health and behavior<sup>1–6</sup>. Here we use these notes for a different purpose: to make inferences about physicians, specifically their level of fatigue when writing the note. We draw on a long tradition of research in natural language processing that links subtle textual cues to an author's psychology<sup>7–9</sup>.

We focus on physicians working in the emergency department (ED), who make critical life and death decisions over shifts lasting between 8 and 12 hours. The work is psychologically and physically demanding, with 65% of physicians reporting burn-out, the highest rate of any medical specialty<sup>10</sup>. A key feature of physician scheduling in this setting is that it is shift-based and variable. Shifts are often clustered on consecutive days, with multi-day gaps in between. So a patient arriving at the ED on any given day might see a physician who has had either a heavy or light workload over the prior week. At the hospital we study, for example, 15% have worked 5 shifts or more in the prior 7 days, while 19% are working their first shift in 7 days.

Given this observation, we propose a novel measure of physician fatigue, based on the notes they write. We obtained all patient notes

from hospital ED visits over a multi-year period, and predicted the number of days a physician had worked over the prior 7 days using the full text of notes written on a given day. Intuitively, if a statistical model can predict last week's workload using only the text of notes written today, that workload must have led to some measurable, lingering effect. Our core assumption is that this effect is mediated through fatigue.

Measuring fatigue through clinical notes offers a novel methodological approach to studying an otherwise difficult-to-quantify phenomenon. Studies have consistently found effects of fatigue on physician behavior, but little impact on patient outcomes<sup>11–16</sup>. These studies, however, study the effect of physician exposure to drivers of fatigue, like workload or circadian factors. But these may be imperfectly correlated to actual physician fatigue at the patient encounter: a physician working a string of overnight shifts is not necessarily fatigued, and a physician working her first daytime shift in a week may have slept poorly the night before. By contrast, our note-based model characterizes physicians' revealed fatigue state at the time of an individual patient encounter. This fine-grained metric lets us study how fatigue affects patient outcomes, on one important measure: physicians' ability to diagnose heart attack.

<sup>1</sup>University of Chicago, Chicago, IL, USA. <sup>2</sup>University of California, Berkeley, CA, USA.  e-mail: [chenhao@uchicago.edu](mailto:chenhao@uchicago.edu)

Beyond clinical care, our results have an important implication for large language models (LLMs), which are rapidly being deployed into patient care and provider workflows. Like text written by physicians, text generated by LLMs can also be analyzed through the lens of predicted fatigue. Our analysis of how predicted fatigue differs between notes written by LLMs vs. those written by physicians raises important questions about the quality of LLM-generated notes.

## Results

### Analytic Strategy

Our analysis uses data on 129,228 consecutive ED encounters with physician notes from a single academic medical center over 2010–2012. This dataset includes patient demographics, the reason for their visit to the ED (the ‘chief complaint’), and key outcomes related to an important physician decision: whether or not the patient is tested for heart attack (via stress testing or catheterization), and the outcome of testing (whether a heart attack was diagnosed and treated, via stenting or open-heart surgery; further details are in<sup>17</sup>).

We identify the attending physician who wrote the clinical note for each of these visits and was thus responsible for the medical decision-making. In total we observe 60 emergency physicians working 11,592 shifts. A shift is defined as consecutive notes entered by the same physician, with each note timestamped within 3 hours of the prior note’s timestamp (the time between the end of a scheduled shift and the start of the next shift is at least 15 hours apart). While in some settings, physicians may write their notes several days after the patient encounter, in our setting, we verify that 99.98% of notes were written on the same day as the patient encounter. We calculate a physician’s workload by counting the number of days worked over a rolling seven-day window ending with the current shift (see Fig. 1). The median is 3, and the mean is 2.88 days worked over the prior week. We define ‘high-workload’ physicians as those who worked at least 4 days prior to the current shift (14.8%), and compare them to ‘low-workload’ physicians, those who are working their first shift in 7 days (19.0%).

Our analytic strategy rests on the assumption that the text of notes written today should be statistically uncorrelated with how many days the physician has worked over the past week, except via the direct effect of prior workload on present physician fatigue state while writing the note. We use the term fatigue as a convenient shorthand to group broadly related factors, including reduced cognitive engagement or attention.

This assumption could be violated if patient characteristics on a given shift differ as a function of prior workload for any reason: In this case, differences in note text could be due to patient factors, not physician fatigue. For example, physicians may be assigned less challenging shifts after a high-workload period, with less acute patients; or physicians may choose less complex patients after such periods (physicians have considerable discretion in patient choice in this setting<sup>18</sup>). We thus carefully test for balance in patient characteristics between high- vs. low-workload physicians. Concretely, we regress patient demographics, patient severity, and chief complaints against the high- and low-workload indicator in the prior week, controlling for time of day, day of week, week of year, year, physician fixed effects. We find no significant differences in demographics and severity measured by patient length of stay; we also find that only 10 of 154 chief complaints (6.5%) have significant differences at the  $p < 0.05$  level (see Supplementary Materials). While it is impossible to verify balance on unmeasured characteristics, this degree of balance on measured factors is reassuring that patients are as-good-as-randomly assigned to high- vs. low-workload physicians.

We set up our prediction model as a binary classification task, to distinguish whether a note is written by a high-workload physician or a low-workload one. To train and evaluate the model, we create a dataset restricted to high- and low-workload physicians, containing 44,556

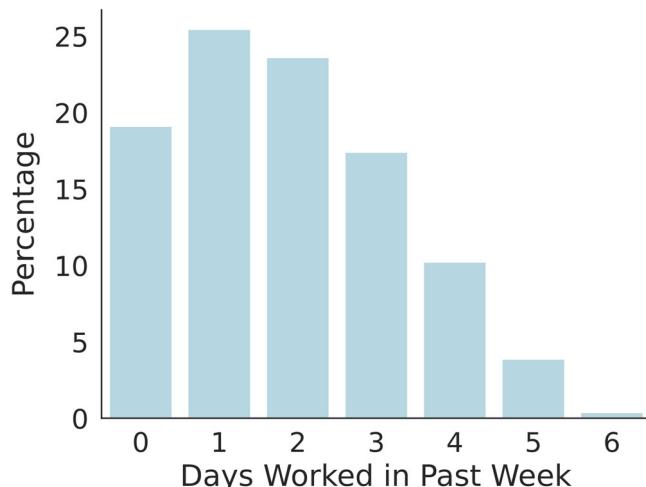
notes from patient encounters evenly balanced (50%–50%) between encounters with high- and low-workload physicians. We randomly divide this at the patient level (so that all visits by a given patient are grouped together) into a training set of 32,784 encounters and a held-out set of 11,772 encounters (see Methods for an overview of the sample and data splits).

Our model uses four categories of interpretable features to classify notes. First, we measure note length (number of words). Second, we create a measure of note predictability by fine-tuning a large language model (GPT-2) on the training set<sup>19</sup>. Specifically, we use *perplexity* which captures the average log likelihood of words being generated from a language model. Intuitively, large language models are trained to predict the next word, and if they can reliably predict what the next word is, it suggests that the note is very predictable. See a detailed explanation in Methods. Third, we measure note readability with Flesch-Kincaid grade<sup>20</sup>. Finally, we quantify the fraction of words in each note according to specific categories: stopwords (e.g., “the”, “is”, “and”); Linguistic Inquiry and Word Count (LIWC) lexicons (e.g., pronouns, affect, and cognitive words<sup>21</sup>); and medical concepts from the SNOMED-CT Ontology<sup>22</sup>. We then train a logistic regression model that uses these features to distinguish between notes written by high- vs. low-workload physicians. We chose to use well-known features from the literature, combined in a simple linear model, because this strategy offers a transparent and interpretable approach to prediction. We also note that, in preliminary experiments, logistic regression had competitive performance compared to more sophisticated models (neural networks and boosted trees).

We emphasize an important, but counterintuitive, aspect of our analytic strategy: while our model is trained to predict a physician’s prior workload—our goal is explicitly not to perfectly predict this variable. Instead, we wish to learn a general model of how fatigue affects the text of notes, by training the model to distinguish notes written by physicians we believe to be more (high-workload) vs. less (low-workload) fatigued. In fact, we believe our model predictions are a better measure of a physician’s ‘true’ fatigue when writing a given note than the actual training label (prior workload) itself, assuming that the only mechanism that the number of days worked in the past seven days (Y) can affect the notes (X) is through fatigue. We provide the intuition for this argument here, and further explanation and a formal proof in Methods; we also provide a range of empirical tests in the following results.

Our model is trained to predict workload from note text, and a ‘perfect’ model would simply predict this variable for all patients seen on a shift (because prior workload does not vary within a shift). However, workload is only one factor contributing to a physician’s latent fatigue state when writing a note. Other idiosyncratic shocks, such as sleep quality, emotional stress, or the intensity of the current shift, may also affect fatigue independently of prior workload. If fatigue affects notes in the same way—whether it is caused by heavy workload, idiosyncratically poor sleep the night before, etc.—a model that learns to predict workload will learn something about the general way fatigue affects note text. As a consequence, its predictions will also be correlated to the idiosyncratic shocks to fatigue (even if these are uncorrelated to workload). Intuitively, the model learns about the effect of fatigue on notes by looking at workload. It may identify signs of fatigue even if the physician does not have a high workload (e.g., the physician is sleep-deprived). So the model may be wrong in useful ways: ‘errors’ (when evaluated against the original label, workload) might mean model predictions are closer to the ‘true’ fatigue state than the actual training label.

To test whether these errors are signal or noise, we next compare predictions to a range of other measures of fatigue, to which the model has not been exposed in training. If predictions correlate to these other measures, it supports the hypothesis that the model has learned general patterns about how fatigue affects notes, which transfer into



**Fig. 1 | Physician work patterns prior to patient encounters.** The histogram shows the percentage of patient encounters by the number of days physicians worked in the seven days prior to writing a note (including the day when the patient encounter happened; 0 means that the patient encounter happened on the first day that the physician worked in the seven-day window) ( $n=129,228$ ).

other settings. Further below, we will compare predictions to important physician decisions that may also be affected by fatigue.

#### Model-predicted fatigue is correlated with independent measures of fatigue

Under the null hypothesis, prior physician workload should be uncorrelated with notes written on a given day, and our model should thus perform no better than chance at predicting workload. Our first test thus simply compares model predictions on notes written by high- vs. low-workload physicians in our balanced hold-out set. To measure whether the model can distinguish the notes, we use AUC, which would be 50% under random guessing. Our model achieves an AUC-ROC of 60.1% (bootstrapped 95% CI: 60.06% - 60.30%), suggesting that there is indeed a statistical correlation between prior workload and note text. An alternative approach is by regressing high workload (as a binary indicator) on model predictions, which shows a large and highly significant coefficient (Table 1, Column 1), controlling for patient characteristics, physician fixed effects, and time controls (time of day, day of week, week of year, and year).

A correlation between note text and prior workload does not identify fatigue as the mechanism. To generate more robust evidence that the model predictions reflect fatigue, we correlate model predictions to several additional measures of physician fatigue. Importantly, the model has not been trained on any of these variables. So these regressions test whether the model has learned generalizable patterns about how fatigue affects note text, based on distinguishing notes written by high vs. low workload physicians. We test whether model predictions correlate to circadian disruption in two ways. First, we compare patients seen by the overnight physician to other patients. To do so, we create an indicator variable for whether the patient arrived between 1:00 a.m. and 5:59 a.m., when only the overnight physician was seeing new patients. In a regression adjusting for time, patient, and physician effects, Table 1, Column 2, shows that model predictions are a significant predictor of whether a note was written on the overnight shift.

Figure 2 shows model predictions graphically as a function of patient arrival time. Consistent with the regression results, model predictions are greater on average for patients arriving between 1 a.m. and 6 a.m.; in fact, predictions increase over that time period, compared to stable model predictions over the rest of the day.

Our second measure of circadian disruption is the variance of physician shift start times over the prior week (including the current shift). The intuition is that a physician who has worked the prior three shifts starting at 7:00 a.m., and is now starting again at 7:00 a.m., is less disrupted than a physician who worked prior three shifts starting at 7:00 a.m., 5:00 p.m., and 11:00 p.m.. Table 1, Column 3 again shows a significant coefficient on model predictions.

Finally, we link model predictions to increasing fatigue over the course of a shift, measured by how many patients a physician has already seen on that shift (Table 1, Column 4), again showing a large and significant coefficient.

These findings strengthen the interpretation that model predictions reflect physician fatigue. While predicted prior workload, in principle, could pick up on spurious correlations or other aspects of physician behavior besides fatigue, the fact that it correlates with several other more specific settings where fatigue is common makes it more likely that we are indeed measuring fatigue and not some other phenomenon.

**External validation.** External validation of any work involving physician notes is challenging because publicly accessible note datasets are rare: free text is difficult to fully de-identify. The few datasets available online, such as i2b2<sup>23</sup> or MIMIC-III<sup>24</sup>, lack much of the fine-grained information on physician shift assignment, timing of shifts, and workload, and fatigue-relevant information we use to validate the model (e.g., physician identifiers). MIMIC-III contains notes from patients admitted to the intensive care unit of one hospital, unlike the sample for our main analysis, which contains notes from emergency department patients. A key problem is that dates are jittered to protect patient privacy. This means we cannot replicate our analysis on variables such as “variation of shift time” and “patients seen prior” in Table 1. However, because MIMIC-III notes are timestamped, we can apply the fatigue prediction model from our main dataset to ascertain whether it correlates to whether a note was written on an overnight shift.

We first restrict to the “admission note”, a summary of the patient’s reason for ICU admission that is the most similar in structure to the emergency department notes from our main sample; we then drop the notes that were written more than 6 hours after the patient was actually admitted, as the physician’s fatigue state would not necessarily correlate to the time of arrival. As in our main analysis, we begin by fine-tuning a language model (GPT-2) in MIMIC-III notes, splitting physician notes by patient following<sup>25</sup> (see “Methods” for details). This allows us to measure physician note perplexity on the test dataset, which we then combine with the other features in our model to generate final predictions on fatigue.

We then define overnight shifts in the same way as in our main sample (patients arriving between 1 am and 6 am), and regress this variable on predicted fatigue, controlling for time of day, physician, patient demographics, and diagnosis. Table 1, Column 5 shows a significant positive relationship of predicted fatigue with the overnight shift indicator ( $p$  value = 0.024). Predicted fatigue score is slightly lower in MIMIC III on average, and the score difference (in SD units) between overnight and the rest of the day is larger in MIMIC (19%) than in our dataset (7%), hence the greater coefficient.

#### Increased model-predicted fatigue is indicative of worse physician decision-making

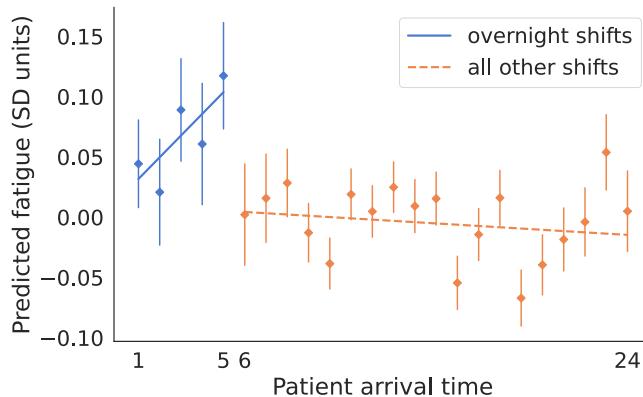
A common dilemma in studies of physician fatigue is the inability to directly measure it. Instead, they measure physician exposure to drivers of fatigue like workload or circadian factors. However, these measures are an imperfect proxy for physician fatigue, due to a range of idiosyncratic factors such as sleep quality, diet, and exercise. As a result, studies relying on them may suffer from attenuation bias and

**Table 1 | Relationship of model predictions to multiple measures of fatigue**

	(1) Hospital Dataset	(2)	(3)	(4)	(5) Ext. Validation
	Workload (Day)	Overnight Shift	Variation in Shift - Start-Time	Patients Seen Prior	Overnight Shift
Fatigue	0.384 <sup>***</sup> (0.087)	0.041 <sup>*</sup> (0.018)	5.606 <sup>*</sup> (2.265)	1.073 <sup>***</sup> (0.298)	0.364 <sup>*</sup> (0.162)
Intercept	3.072 <sup>***</sup> (0.102)	0.158 <sup>***</sup> (0.019)	29.855 <sup>***</sup> (2.669)	9.011 <sup>***</sup> (0.352)	-0.239 (0.288)
Controls					
Time of day	YES	NO	YES	YES	YES
Day of week	YES	YES	YES	YES	N/A
Week of year	YES	YES	YES	YES	N/A
Year	YES	YES	YES	YES	N/A
Demographics	YES	YES	YES	YES	YES
Chief complaint	YES	YES	YES	YES	YES
Physician	YES	YES	YES	YES	YES
Length of stay	YES	YES	YES	YES	YES
Insurance type	YES	YES	YES	YES	YES

Columns 1-4 show results from an internal hold-out set ( $n = 34,175$ ); Column 5 shows results from a wholly independent dataset, MIMIC-III ( $n = 1216$ ). Column 1: regression of physician workload over the prior week (the model's training label) on model predictions. Columns 2-4: regressions of several other measures of fatigue (which the model has never seen) on model predictions, including an indicator for whether the note was written on the overnight shift (Column 2); the degree of circadian disruption over the prior week, as measured by the variance of a physician's shift start times over that week (Column 3); and increasing patient volume, measured by the number of patients the physician has seen on-shift before writing a given note (Column 4). Column 5 replicates Column 2 in the MIMIC-III dataset of physician notes from another hospital's intensive care unit, showing that model predictions generalize beyond the training dataset.  $p$ -values of fatigue coefficients are 9.86e-6, 0.02, 0.013, 0.0003, 0.024 for Columns 1-5, respectively, based on two-sided tests.

$p < 0.05$ ,  $p < 0.01$ ,  $p < 0.001$ .



**Fig. 2 | Predicted fatigue vs. patient arrival time.** The y-axis shows model predictions in standard deviation units. The x-axis shows patient arrival time. Patients arriving after 1 a.m. and before 6 a.m. (blue,  $n = 2595$ ) can only be seen by the overnight physician, while other patients can be seen by any of the multiple physicians working at a given time (yellow,  $n = 31,580$ ). The center points show the average predicted fatigue for patients arriving in that hour, and the error bars show standard error.

fail to find an effect. As a result, most studies to date have found effects of (drivers of) fatigue on physician behavior, but little impact on patient outcomes<sup>11–16</sup>.

We hypothesize that our measure of fatigue might address some of these shortcomings. Our measure is calculated at the patient-note level, and importantly—as noted before and in Methods—is a more accurate measure of “revealed” physician fatigue than inputs like workload, circadian disruption, etc.

To test this, we use a measure developed in prior work in this same dataset<sup>17</sup>, examining the quality of a critical decision in the emergency setting: whether or not to test a patient for acute coronary syndromes (ACS, colloquially: heart attack). This decision is an important yet

challenging one. ACS is important to diagnose in a timely manner, but testing for it is invasive and resource-intensive. Research has shown widespread over-testing: some patients have such predictably low likelihood of testing positive that performing the test offers limited clinical value, as measured by the yield of testing, i.e., the fraction of tests that come back positive<sup>26,27</sup>. Following this literature, we view yield as a patient-centered measure of the quality of physician decision-making: a higher yield means a higher rate of diagnosing ACS and unlocking substantial health benefits for the patient, while a lower yield means incurring the costs and risks of testing with no clear patient benefit.

Table 2 presents the results of this analysis. We first regress the yield of testing on the number of days worked in the prior week, controlling for time, patient, and physician effects. This analysis simulates the standard way of testing the effect of physician fatigue on patient outcomes: measuring the relationship between a patient outcome and a coarse measure of physician workload that does not vary at the patient level. Similar to prior work on physician fatigue, we find no effect: the coefficient on days worked is small and insignificant. We then repeat the analysis, substituting our note-based measure of fatigue at the patient level instead of the coarse measure of days worked. Here, the coefficient is large and significant: for each one standard deviation increase in fatigue, the yield of testing decreases by 19.0%. This result indicates that fine-grained measures of fatigue, like the one we use here, are a promising way to measure and elucidate the consequences of physician fatigue.

#### Note predictability is an important predictor of fatigue

To provide some intuition on the features of notes used by our model to predict fatigue, we show correlations of model features with physician workload (specified as an indicator of high- vs. low-workload, the label used for training the model) in Table 3 (See the Supplementary Materials for complete results).

The most highly correlated feature is the predictability of a note, as measured by note perplexity. Intuitively, this feature measures how hard it is for a large language model to predict a given word in a note,

**Table 2 | Effect of workload and effect of predicted fatigue on yield of testing for heart attack (more precisely, the positive rate of tests done in the few days after the ED visit) (n = 1017)**

	Effect of workload on yield of testing	Effect of predicted fatigue on yield of testing
Coefficient	-0.0081 (0.009)	-0.2935* (0.147)
Controls		
Time of day	YES	YES
Day of week	YES	YES
Week of year	YES	YES
Year	YES	YES
Demographics	YES	YES
Chief complaint	YES	YES
Physician	YES	YES
Length of stay	YES	YES
Insurance type	YES	YES

Column 1 shows a regression of testing yield on a physician's prior week workload, which is not significant ( $p = 0.383$  in a two-sided test). Column 2 shows a regression of testing yield on the predicted fatigue score, calculated on the basis of the patient's note, which is significant and negative ( $p = 0.047$  in a two-sided test).

$p < 0.05^*$ ,  $p < 0.01^{**}$ ,  $p < 0.001^{***}$ .

**Table 3 | Features used by the model: Pearson's correlation between physician workload with the high- vs. low-workload indicator (n = 43,730), as measured by workload over the prior week**

Feature	Correlation with physician workload	P value
Word Un-predictability		
Perplexity (log)	-0.092**	5.34e-83
Cognitive words (LIWC)		
Insight (fraction)	-0.09**	4.34e-75
Certainty (fraction)	0.075**	1.07e-56
Pronouns (LIWC)		
First person singular pronouns (fraction)	-0.07**	7.85e-48
Impersonal pronouns (fraction)	-0.049**	3.82e-22
Affect (LIWC)		
Anger (fraction)	0.025**	1.86e-07
Readability		
Flesch-Kincaid grade	-0.05**	4.88e-24

More details are in the Supplementary Materials.  $p$  values are derived from a two-sided test.  
 $p < 0.05^*$ ,  $p < 0.01^{**}$ ,  $p < 0.001^{***}$ .

conditional on the preceding words in the note. We find that perplexity decreases with workload, implying that fatigued physicians tend to write more predictable notes. This suggests that fatigued physicians produce more formulaic or generic documentation, potentially reflecting reduced cognitive engagement. Related to predictability, the readability of notes, based on the complexity of sentences and words, is lower when physicians are fatigued.

Fractions of words related to insight and certainty, as categorized by the standard LIWC corpus<sup>21</sup>, are linked to model predictions. The insight category includes words such as "believe," "reveal," "think," and "explain," and the certainty category includes words such as "clear," "certain," "apparent," and "never." Fatigued physicians tend to use fewer words in the insight category and more words in the certainty category. In a related pattern, we also find that fatigued physicians are less likely to use first-person singular pronouns: many insight words

typically appear with first-person singular pronouns such as "I believe." Other work has linked similar patterns with withdrawal and less personal agency<sup>28</sup>. Finally, we find that the fraction of words related to anger, for example, "lying", "assault", and "threat", is also positively correlated with fatigue. Recall that we find no correlation between patients' reasons for visiting the ED (their chief complaint) and physician workload, suggesting that differential assignment of assault cases to physicians does not explain this finding.

## Discussion

We demonstrated the potential of detecting physician fatigue using the notes that they wrote. Our predicted fatigue allowed us to reveal connections with decision quality. Our finding highlights the role of clinical notes not only as a medium of storing medical information but also as a window into physician decision-making.

Our results have important implications for the emerging literature on LLMs trained on medical notes<sup>3</sup>. First, while LLMs excel at many tasks related to language, they were substantially worse than a supervised model at distinguishing notes written by high vs. low workload physicians. Using Vicuna-7B<sup>29</sup> for zero-shot classification, the AUC-ROC was 53.2% (95% CI: 53.1–53.3), significantly lower than our approach ( $p < 0.01$ ; see the Supplementary Materials). This is perhaps unsurprising as fatigue is a latent, rather than an explicit, component of physician notes, and there are few training data to learn from; it also illustrates that specialized models still have many advantages for medical tasks<sup>30</sup>.

Second, as shown above, a key feature linked to fatigue in our model was the predictability of the next word in a note. Text generated by LLMs is, by construction, predictable: next-word prediction is the basis for the core training task around which LLMs are built. This raises the possibility that LLM-generated notes may be characterized by many of the same features that make for fatigued-appearing physician notes. To test this, we prompted an LLM with the first sentence of a real note and asked it to complete the patient history. We then compared this to the real patient history and quantified measured fatigue for each pair. Patient histories generated by LLMs had 74% higher predicted fatigue than genuine physician notes, and in particular had higher scores on predictability and fraction of anger words (see the Supplementary Materials). This illustrates some of the dangers of relying on LLMs as a substitute for physicians in generating clinical notes: If fatigue-related features are indicative of reduced note quality, this finding suggests that LLM-generated notes may encode lower-quality clinical information than initially apparent.

Furthermore, as writing is an important form of thinking<sup>31</sup>, automation may cause physicians to skip some of the active thinking components of the patient encounter. Other recent work also shows that LLMs can perpetuate racial biases in generated medical texts<sup>32,33</sup>. It calls into question the strategy of relying on humans to evaluate the quality of text produced by LLMs, which is pervasive in the current literature<sup>34,35</sup>. While we were not able to explore this in the current work, we suspect clinicians may be just as unable to identify fatigued notes as LLMs are. And given the possibility that fatigued notes miss important details—as suggested by the correlation between note-based fatigue and low-yield testing for heart attack—new benchmarks are urgently needed to assure quality before LLMs become widely deployed.

Last but not least, instead of the straightforward yet potentially harmful application of automating note-writing<sup>33</sup>, LLMs provide opportunities to transform the information solicitation process in physician-patient encounters. The predictable information in the notes may not require precious physician-patient interaction to acquire, while LLMs may help physicians identify valuable information to solicit by suggesting questions that lead to information with low predictability. While rested physicians may write notes that are less predictable, which may correlate with lower readability, LLMs may

provide effective ways to decouple the process of meeting patients and writing notes from the process of reading notes by simplifying the language when presented to patients or other physicians.

In short, we advocate responsible adoption of LLMs as a writing assistance tool in the healthcare domain. It is important to ensure that LLMs serve as an augmentation tool to improve the note-taking practice, without eroding the agency of physicians and reducing the value of information in clinical notes.

## Methods

Our study complies with all relevant ethical regulations and was approved by the Mass General Brigham IRB (Protocol 2018P000907), which provided a waiver of consent for use of administratively-collected patient data, including notes. No data left the hospital's computing environment, including data fed into language models, as these were implemented inside the same environment.

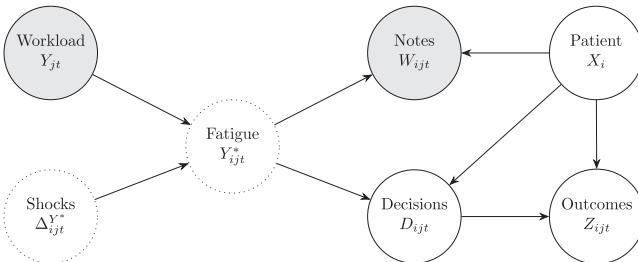
### Behavioral model for the prediction task

Here we provide a more detailed argument for why our predictions, despite being trained on prior workload, are a better measure of a physician's true fatigue when writing a given note than prior workload itself.

Our model of the causal relationships among the variables we consider is shown in Fig. 3, where dotted nodes are unobserved, and shaded nodes are components of our statistical model. In this model, doctor  $j$  sees patient  $i$  with characteristics  $X_i$  during shift  $t$ . We assume doctor  $j$ 's past-week workload, denoted as  $Y_{jt}$ , affects unmeasured "true" fatigue when treating patient  $i$ ,  $Y_{ijt}^*$ ,

$$Y_{ijt}^* = Y_{jt} + \Delta_{ijt}^Y$$

where  $\Delta_{ijt}^Y$  is a set of idiosyncratic random shocks and coping strategies affecting true fatigue when the physician sees patient  $i$ . We assume (and verify)  $Y \perp X$ . We also assume that  $Y_{jt}$  affects the written note  $W_{ijt}$



**Fig. 3 | Variable causal relationships.** Dotted nodes are unobserved, shaded nodes are components of our statistical model  $Y_{jt} = g(W_{jt})$ .

**Table 4 | Summary of basic statistics**

Category	All	Training set	Heldout set
Encounters	129,228	95,053	34,175
Patients	77,854	57,311	20,543
Physicians	60	60	60
Demographics			
Age	41.8	41.9	41.8
White	0.50	0.50	0.50
Black	0.22	0.22	0.21
Hispanic	0.18	0.18	0.18
Female	0.60	0.60	0.60
Male	0.40	0.40	0.40

and physician decision quality  $D_{ijt}$  only via  $Y$ . Decisions  $D_{ijt}$ , and patient factors  $X_i$  jointly determine patient outcomes  $Z_{ijt}$  at that visit.

Intuitively, because our statistical model takes the form  $Y_{jt} = g(W_{jt})$  and the label  $Y_{jt}$  is the same for all patients  $i$  seen by doctor  $j$  on shift  $t$ , it would seem that following a "perfect" model would predict a constant  $\hat{Y}_{ijt} = E[\hat{Y}_{ijt}]$ . However, note that both  $Y_{jt}$  and  $\Delta_{ijt}^Y$  affect  $Y_{ijt}^*$ , and there are no other paths by which either can affect notes except via  $Y_{ijt}^*$ . As a result,  $g(\cdot)$  learns about features of notes  $W$  that predict workload  $Y$ ; but because this relationship is mediated through  $Y$ , it also learns about the idiosyncratic shocks  $\Delta^Y$  that affect  $W$  through the same channel  $Y$  (and it does so even though  $Y \perp \Delta^Y$ , by construction).

This has a key implication for a model that uses  $W_{ijt}$  to predict  $Y_{jt}$ : it will *not* predict  $Y_{jt}$  accurately because  $W_{ijt}$  is correlated with  $\Delta_{ijt}^Y$ , while  $Y_{jt}$  is not. Notes can be written

$$W_{ijt} = Y_{ijt}^* + X_i = (Y_{jt} + \Delta_{ijt}^Y) + X_i$$

which means that predictions

$$\hat{Y}_{ijt} \approx E[Y_{ijt} | W_{ijt}] = E[Y_{ijt} | Y_{jt}, \Delta_{ijt}^Y, X_i] = E[Y_{ijt} | Y_{jt}, \Delta_{ijt}^Y]$$

where  $X_i$  is dropped because (as we show)  $Y \perp X$ . As a result,  $\hat{Y}_{ijt}$  will contain "errors" when evaluated against the original label  $Y_{jt}$ , because of idiosyncratic variations in true fatigue  $\Delta_{ijt}^Y$ . Indeed, as the influence of  $\Delta_{ijt}^Y$  on  $Y_{ijt}^*$  grows large relative to  $Y_{jt}$ , our predictions will approach closer to  $\hat{Y}_{ijt}$ , but contain larger "errors" relative to the "true" label  $Y_{jt}$ .

A formal proof of the statement in a simple uni-dimensional linear setting follows. Assume  $W_{ijt} = (Y_{jt} + \Delta_{ijt}^Y)A$ , then, the regression coefficient of  $Y \sim W$  is (we omit the subscripts for ease of notation):

$$\begin{aligned} \hat{\beta} &= (W^T W)^{-1} W^T Y \\ &= (A^T (Y + \Delta^Y)^T (Y + \Delta^Y) A)^{-1} (A^T (Y + \Delta^Y)^T) Y \\ &= \frac{1}{A} \cdot \frac{1}{Y^T Y + (\Delta^Y)^T \Delta^Y} \cdot (Y^T Y). \end{aligned}$$

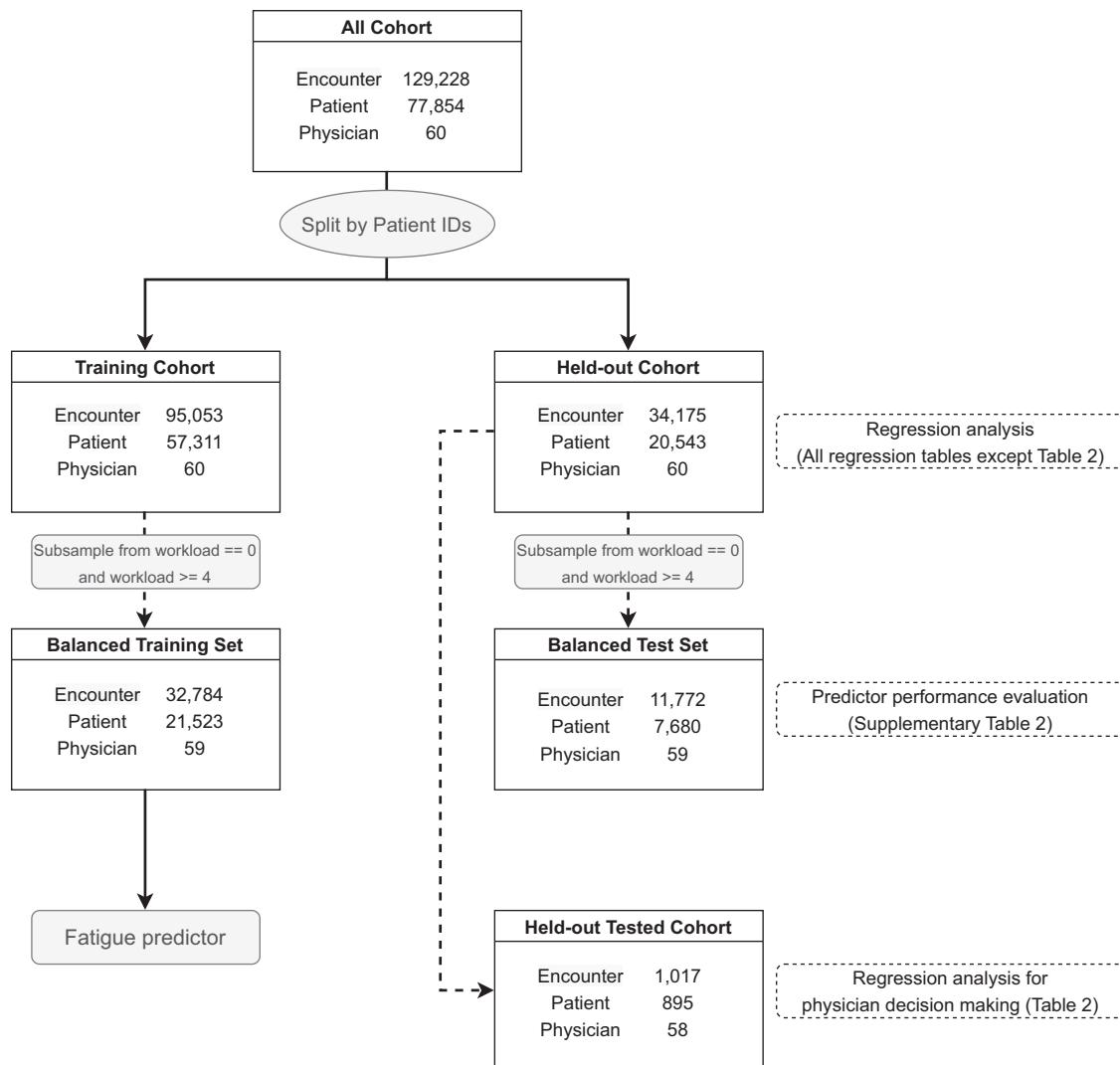
Recall  $A$  is a scalar in this case. It follows

$$\begin{aligned} \hat{Y} &= W \hat{\beta} \\ &= (Y + \Delta^Y) A \cdot \frac{1}{A} \cdot \frac{1}{Y^T Y + (\Delta^Y)^T \Delta^Y} \cdot (Y^T Y) \\ &= \frac{Y^T Y}{Y^T Y + (\Delta^Y)^T \Delta^Y} (Y + \Delta^Y) \end{aligned}$$

We observe that  $\hat{Y}$  is a shrunk version of  $Y + \Delta^Y$ , and how well it captures  $Y$  is determined by the relative magnitude of  $Y$  and  $\Delta^Y$ .

As a result, our model predictions could be "wrong" (with respect to  $Y_{jt}$ ) because they are good—i.e., closer to  $Y_{ijt}^*$ . Alternatively, the model could be wrong in less useful ways—it could simply be making bad predictions. Importantly, the graph above provides an empirical test of which is more true: good predictions will correlate with patient outcomes  $Z_{ijt}$ , via  $Y_{ijt}^*$ , which we test in the relationship of measured fatigue to physician decision making.

Lastly, this model helps us contextualize the fact that patient characteristics  $X_i$  may correlate with  $\hat{Y}_{ijt}$  but not  $Y_{jt}$ . In our model, this can only be because of correlations between  $X_i$  and the idiosyncratic shocks to fatigue  $\Delta_{ijt}^Y$  that can vary at the patient level and are incorporated into  $\hat{Y}_{ijt}$ . We test that these characteristics are not correlated with workload. Failing this, it is possible—as considerable research has demonstrated both inside and outside of the hospital—that latent psychological factors are present at the patient level that mimic fatigue when dealing with demographically different patients.



**Fig. 4 | Data diagram.** The dataset contains 129,228 encounters from 77,854 patients and 60 physicians.

## Dataset

Our physician notes dataset was obtained from Mass General Brigham in the United States from 2010 to 2012. Figure 4 presents the data splitting workflow in this work. Table 4 shows basic statistics of the dataset, and Table 5 shows the top 10 chief complaints in our dataset. The notes collected were written on the same day of the patient encounter, except 32 notes (0.02%) that were written after the encounter date.

**Note format.** Generally, physician notes are semi-structured: they are grouped into sections, and each section includes free texts. The most common sections are History of Present Illness, Physical Examination, and Assessment and Plan. First, the History of Present Illness section describes the reason for the encounter. Then, the Physical Examination section is also included to understand what preliminary tests have been done in this encounter. Finally, the Assessment and Plan section presents the thought processes for the final diagnosis.

**Table 5 | Top 10 frequent chief complaints**

Chief complaint	count
abd pain	35895
chest or esophag pain	22679
dyspnea	16381
back pain	12920
fever chills	11389
fall	10382
elbow wri hand finger cmplt	10266
leg cmplt	9192
foot ankle toe cmplt	9091
headache	8691

**External validation: MIMIC-III dataset.** We split the MIMIC-III physician notes into 89409/29804/22411 notes for training/validation/test dataset based on patient IDs to check generalizability to unseen patients. We fine-tune the language model for computing the perplexity of physician notes as one of the input features of the fatigue predictor. For regression analysis, we only select the admission notes of MIMIC-III physician notes in the test set to match our internal dataset. Feature values for analysis are in the Supplementary Materials. We also experimented with splitting by physician ids, which yields varying results, suggesting that seeing physician notes in the training set is important for the fine-tuning language model for our approach.

## Fatigue predictor

**Features.** For every note used in this paper for the fatigue predictor, we drop the beginning and the end of note, which are generated by the

system. Instead of directly using textual content in the physician note, we select features derived from note content to avoid spurious correlation that can be inferred from the actual text and improve the interpretability of the predictor. We include basic statistics of notes, i.e., note length, fraction of stopwords, fraction of medical words, and readability. Also, we use LIWC<sup>21</sup> lexicon to capture cognitive and affective processes of physicians. Finally, we introduce a novel feature of note predictability obtained from a fine-tuned large language model, GPT-2 with 117 million parameters<sup>19</sup>. See the Supplementary Materials for statistics of feature values.

After extracting the features, we standardize the features to have zero mean and unit variance. We also include the dummy chief complaint features to control for the effect of chief complaints on note-level fatigue predictions.

**Perplexity and language modeling details.** A language model predicts the probability of the next token, and thus assigns probability to a sentence or document:  $p(X) = \prod_{i=1}^n p(x_i|x_1, \dots, x_{i-1})$ , where  $X$  is the document and  $x_i$  denotes  $i$ -th token in the document. Perplexity is then computed as  $\text{perplexity}(X) = 2^{-\frac{1}{n} \log_2 p(X)}$ . A lower perplexity indicates that the text is more predictable by the language model.

Recent studies in large-scale language models show that fine-tuning from the pretrained checkpoint can improve the domain-specific language modeling<sup>36,37</sup>. In this work, we finetune GPT-2 (117M) on our training set and compute the perplexity of each note (document) in the whole dataset. The greater the log perplexity, the lower the note predictability.

Related work using text analysis to differentiate between written descriptions of real vs. imagined events has also found strong correlations with note perplexity<sup>9</sup>.

**Classification experiment setup.** The balanced predictor dataset is derived from the original dataset (see the Supplementary Materials for detailed data split). We define a note as fatigued note if the note is written by a physician working last 4 days before today. In contrast, we define a note as non-fatigued note if the note is written by a physician working 0 days before today in a 7-day span. For each chief complaint, we sample the same number of fatigued and non-fatigued notes from the original training and hold-out set to form the balanced dataset. Given the fatigue dataset, we use the logistic regression model with regularization as our classifier. We conduct hyperparameter search on the regularization term with 5-fold cross validation on the training set.

**Baseline model.** For the baseline model, we only use chief complaint as the input feature. We use the same logistic regression model with regularization as our classifier.

**Prompt for the Vicuna-7B model.** To see how pretrained language models can be used for the fatigue prediction task in a zero-shot manner, we use the Vicuna-7B model to perform the task. We sample 1000 pairs of fatigued and non-fatigued notes from the balanced test set and ask the Vicuna-7B model to predict which note is written by a more fatigued physician with the following prompt.

Prompt:  
 Note 1: {shuffled\_notes[0] }  
 Note 2: {shuffled\_notes[1] }  
 Task:  
 Analyze the above two physician notes and assess  
 which one appears to be written by a more fatigued  
 physician.  
 Answer the question at the end by selecting either [Note  
 1] or [Note 2].  
 Only reply the answer.  
 Do not include any other information.

## Regression setup

We present details on regression analysis.

## Control variables.

1. Time control includes time of day, day of week, week of year, and year as categorical variables.
2. Chief complaint control includes binarized chief complaints. An encounter can have multiple chief complaints.
3. Physician control uses the physician id as a categorical variable.
4. Patient demographics control uses the patient sex and race as categorical variables and patient age as the numerical variable.
5. Length of stay control represents the severity of patient's current encounter as the numerical variable.
6. Insurance class control is a categorical variable accounting for a patient's socioeconomic status.

Due to the small number of tested observations in the test set for the regression on the yield of testing ( $n=1017$ ), we only control for chief complaints and physicians with at least 20 occurrences.

**Circadian disruption: shift start time variance definition.** We compute the variance of starting time (hour of day) of the shifts in the past week, including the current shift, as one of our circadian disruption measurements. To mitigate the effect of midnight time change leads (23 vs. 0) to a huge variance, we shift the starting time back by 6 hours. For instance, two night shifts originally beginning at 23:00 and 00:00 are 17 and 18 in the adjusted schedule for the sake of computing the variance. Empirically, very few shifts started between 0 and 6 a.m.

**Dataset for regression analysis.** Regressions are done with our test set with 34,175 observations, except the yield of testing regressions, which only includes 1017 observations of the tested cohort in the test set.

## Limitations

A key limitation of our study lies in the dataset that we have access to. It comes from a major hospital in the United States. Although we validated parts of our findings on an external dataset from MIMIC, thorough validations across healthcare systems will further improve our understanding of the relationship between clinical notes and physician decision-making. Another limitation comes from the models that we used. Physician notes contain protected health information that is difficult to fully remove, meaning the data cannot leave the hospital computing environment. Because the resources available inside this environment are limited, we were not able to train or deploy larger models. Finally, writing clinical notes is an evolving practice as doctors adopt AI tools in the process. We encourage future studies to explore the effect of AI assistance in physician decision-making.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Data supporting the findings of this study are available in the article and its Supplementary information. Source data are provided as Source Data file and may be obtained from the corresponding authors upon request. The data used for the primary analysis consist of individual patient records, including free-text physician notes, which are challenging to fully deidentify. As a result, the IRB did not approve public data sharing. External validation was performed using the publicly available MIMIC-III dataset (<https://physionet.org/content/mimiciii/1.4/>). Source data are provided with this paper.

## Code availability

Code that supports the main findings of this study are available on GitHub: <https://github.com/ChicagoHAI/physician-fatigue>.

## References

1. Ghassemi, M. et al. Unfolding physiological state: mortality modelling in intensive care units. In *Proc. KDD* (ACM Digital Library, 2014).
2. Sheikhalishahi, S. et al. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med. Inform.* **7**, e12239 (2019).
3. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 1–6 (2023).
4. Adams, G., Alsentzer, E., Ketenci, M., Zucker, J. & Elhadad, N. What's in a summary? Laying the groundwork for advances in hospital-course summarization. In *Proc. NAACL* (Association for Computational Linguistics, 2021).
5. Zhang, Y. et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nat. Protoc.* **14**, 3426–3444 (2019).
6. Althoff, T., Clark, K. & Leskovec, J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Trans. Assoc. Comput. Linguist.* **4**, 463–476 (2016).
7. Pennebaker, J. W. The secret life of pronouns. *N. Sci.* **211**, 42–45 (2011).
8. Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J. & Potts, C. No country for old members: user lifecycle and linguistic change in online communities. In *Proc. 22nd International Conference on World Wide Web*, 307–318 (2013).
9. Sap, M., Horvitz, E., Choi, Y., Smith, N. A. & Pennebaker, J. W. Recollection versus imagination: exploring human memory and cognition via neural language models. In *Proc. ACL* (Association for Computational Linguistics, 2020).
10. Carbajal, E. 29 physician specialties ranked by 2022 burnout rates. <https://www.beckershospitalreview.com/hospital-physician-relationships/29-physician-specialties-ranked-by-2022-burnout-rates.html> Accessed: 2023-07-25 (2022).
11. Landrigan, C. P. et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *N. Engl. J. Med.* **351**, 1838–1848 (2004).
12. Barger, L. K. et al. Impact of extended-duration shifts on medical errors, adverse events, and attentional failures. *PLoS Med.* **3**, e487 (2006).
13. Lockley, S. W. et al. Effect of reducing interns' weekly work hours on sleep and attentional failures. *N. Engl. J. Med.* **351**, 1829–1837 (2004).
14. Johns, M. M. et al. *Resident Duty Hours: Enhancing Sleep, Supervision, and Safety* (National Academies Press, 2009).
15. Ayas, N. T. et al. Extended work duration and the risk of self-reported percutaneous injuries in interns. *JAMA* **296**, 1055–1062 (2006).
16. Arnedt, J. T., Owens, J., Crouch, M., Stahl, J. & Carskadon, M. A. Neurobehavioral performance of residents after heavy night call vs after alcohol ingestion. *JAMA* **294**, 1025–1033 (2005).
17. Mullanathan, S. & Obermeyer, Z. Diagnosing physician error: a machine learning approach to low-value health care. *Q. J. Econ.* **137**, 679–727 (2022).
18. Chang, C. Y. & Obermeyer, Z. Association of clinical characteristics with variation in emergency physician preferences for patients. *JAMA Netw. open* **3**, e1919607 (2020).
19. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Tech. Rep.* (2019).
20. Kincaid, P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. In *Chief of Naval Technical Training: Naval Air Station Memphis* (1975).
21. Pennebaker, J. W., Francis, M. E. & Booth, R. J. *Linguistic Inquiry and Word Count* (Lawrence Erlbaum Associates, 2001).
22. El-Sappagh, S., Franda, F., Ali, F. & Kwak, K. S. SNOMED CT standard ontology based on the ontology for general medical science. *BMC Med. Inf. Decision Making* **18**, 76 (2018).
23. Stubbs, A., Kotfila, C., Xu, H. & Uzuner, Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *J. Biomed. Inform.* **58 Suppl**, S67–77 (2015).
24. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
25. Harutyunyan, H., Khachatrian, H., Kale, D. C. & Galstyan, A. G. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 96 (2017).
26. Hermann, L. K. et al. Yield of routine provocative cardiac testing among patients in an emergency department-based chest pain unit. *JAMA Intern. Med.* **173**, 1128–1133 (2013).
27. Stripe, B., Rechenmacher, S., Jurewitz, D., Lee, C. & Schaefer, S. The diagnostic yield of cardiac catheterization in low-risk troponinemia. *JAMA Intern. Med.* **173**, 2088–2090 (2013).
28. Konopasky, A., Durning, S. J., Artino, A. R., Ramani, D. & Battista, A. The linguistic effects of context specificity: exploring affect, cognitive processing, and agency in physicians' think-aloud reflections. *Diagnosis* **7**, 273–280 (2020).
29. Chiang, W.-L. et al. Vicuna: an open-source chatbot impressing gpt-4 with 90%\* chatgpt quality (2023).
30. Lehman, E. et al. Do we still need clinical language models? In *Proc. CHIL* (Proceedings of Machine Learning Research, 2023).
31. Menary, R. Writing as thinking. *Lang. Sci.* **29**, 621–632 (2007).
32. Adam, H. et al. Write it like you see it: detectable differences in clinical notes by race lead to differential model recommendations. In *Proc. AIES* (2022).
33. Zack, T. et al. Coding inequity: assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. *Lancet Digit. Health* **6**, e12–e22 (2024).
34. Nayak, A. et al. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Int. Med.* **183**, 1026–1027 (2023).
35. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
36. Gururangan, S. et al. Don't stop pretraining: adapt language models to domains and tasks. In *Proc. of ACL* (Association for Computational Linguistics, 2020).
37. Wang, Y. et al. Domain-specific pretraining for vertical search: case study on biomedical literature. In *Proc. KDD* (Association for Computing Machinery, 2021).

## Acknowledgements

This work is in part supported by IIS-2126602 (C.H. and C.T.) and a Sloan Research Fellowship (C.T.).

## Author contributions

C.H., Z.O., and C.T. designed the research. Z.O. obtained the main dataset. C.H. implemented the experiments and performed all analyses. C.H., Z.O., and C.T. wrote the paper.

## Competing interests

C.H. has no competing interests to disclose. Unrelated to the manuscript, Z.O. holds equity in Dandelion Health, a for-profit company in the health data space. Unrelated to the manuscript, C.T. is also employed by Abridge AI, a for-profit company in healthcare.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-60865-4>.

**Correspondence** and requests for materials should be addressed to Chenhao Tan.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025