

# Supplementary Materials for Clinical Notes Reveal Physician Fatigue

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

<sup>1</sup>University of Chicago.

<sup>2</sup>University of California, Berkeley.

## 1 Detailed Information about the Fatigue Predictor

### Feature Statistics

Supplementary Table 1 shows basic statistics for different features.

### Model Performance

Supplementary Table 2 shows detailed performance for the prediction task on the balanced hold-out test set. Our model with note features outperforms the baseline model with only chief complaint as the input feature.

### Model Coefficients

Supplementary Table 3 shows the logistic regression model coefficients excluding chief complaints. Note predictability has the highest coefficient of -0.271, the same as the feature correlation analysis. Note that the input features are standardized.

### Additional Discussion on Features of Notes that Correlate with Fatigue

In addition to features mentioned in the main paper, we also observe a negative correlation between workload and note length, indicating doctors write shorter notes when they have a high workload (days worked in the past week). The doctors with a higher workload also write notes with a higher fraction of “inhibition” and “exclusive” words in the LIWC cognitive category, where

	mean	standard error
<b>Note characteristics</b>		
log perplexity	1.1969	0.0008
note length	328.0966	0.4137
fraction of stopwords	0.4027	0.0002
fraction of medical words	0.4815	0.0001
<b>LIWC Pronoun</b>		
fraction of pronoun	0.0755	0.0001
fraction of first person singular pronouns	0.0030	0.0000
fraction of first person plural pronouns	0.0024	0.0000
fraction of second person pronouns	0.0000	0.0000
fraction of third person singular pronouns	0.0463	0.0001
fraction of third person plural pronouns	0.0005	0.0000
fraction of impersonal pronouns	0.0229	0.0000
<b>LIWC Affect</b>		
fraction of affective processes	0.0360	0.0000
fraction of positive emotions	0.0151	0.0000
fraction of negative emotions	0.0204	0.0000
fraction of anxiety	0.0055	0.0000
fraction of anger	0.0009	0.0000
fraction of sadness	0.0029	0.0000
<b>LIWC Cognitive</b>		
fraction of cognitive processes	0.1260	0.0001
fraction of insight	0.0154	0.0000
fraction of causation	0.0049	0.0000
fraction of discrepancy	0.0064	0.0000
fraction of tentative	0.0221	0.0000
fraction of certainty	0.0047	0.0000
fraction of inhibition	0.0056	0.0000
fraction of inclusive	0.0581	0.0000
fraction of exclusive	0.0196	0.0000
<b>Readability</b>		
Flesch-Kincaid grade	8.1098	0.0037

**Supplementary Table 1:** Feature statistics for the tiredness predictor over the whole dataset.

examples of “inhibition” words include “block” and “constrain”, and example words of “exclusive” include “but”, “except”, and “without”.

	CC Baseline	CC + Note Features	Vicuna-7B Zero-shot*
AUC-ROC	50.0	60.1	53.5
Accuracy	50.1	57.2	53.2
F1	38.9	56.6	49.1

**Supplementary Table 2:** Model performance on fatigue prediction. The Chief Complaint (CC) baseline is at the chance level. The pretrained instruct large language model Vicuna-7B achieves better than random accuracy and AUC-ROC scores in a zero-shot manner. The proposed fatigue predictor with note features outperforms both the baseline model and the large language model. \* Note that Vicuna-7B is tested on 1000 pairs of fatigue and relaxed notes sampled from the balanced test set.

## 2 Robustness Checks for the Fatigue Predictor

### *Workload vs. patient demographics and chief complaints*

We regress patient demographics on our target label, i.e., workload, to ensure no correlation between workload and patient demographics. Supplementary Table 4 shows that workload is not correlated with patient demographics. We then regress chief complaints on workload and show that the correlation with chief complaints is not significant (Supplementary Figure 1). We also regress length of stay, which can be seen as a measure of the severity of the patient, on workload. We found no correlation between length of stay and workload ( $p=0.31$ ).

### *Days worked in the past week vs. patient seen in a day*

We further evaluate the balance of other patient characteristics, such as the number of patients a physician sees in a day, to rule out potential confounding factors in how the number of days worked in the past week affects the physician. Thus, we ran a regression of patients seen in a day against days worked in a week controlled for physician and time, and we found no significant effect between the two variables ( $p=0.673$ ).

### *Variance for patient arriving at the same time*

Supplementary Figure 2 shows that for each patient arrival time there exists substance variance between estimated fatigue from notes.

## 3 Characteristics of Generated Notes

### 3.1 Measured with note features

As shown in Supplementary Table 6, in addition to the perplexity difference mentioned in the main paper, we observed that LLM-generated notes usually have shorter lengths and a lower fraction of positive emotion words. In general, notes generated with the sampling decoding method are more similar

Feature	Coefficient
<b>Note Feature</b>	
note length	-0.04971
log perplexity	-0.271
fraction of stopwords	0.12594
fraction of medical words	0.04612
<b>LIWC Pronoun</b>	
fraction of pronoun	0.06129
fraction of first person singular pronouns	-0.14271
fraction of first person plural pronouns	0.06668
fraction of second person pronouns	-0.01114
fraction of third person singular pronouns	-0.07235
fraction of third person plural pronouns	0.04096
fraction of impersonal pronouns	-0.10904
<b>LIWC Affect</b>	
fraction of affective processes	0.02486
fraction of positive emotions	-0.05055
fraction of negative emotions	-0.08713
fraction of anxiety	0.00322
fraction of anger	0.07171
fraction of sadness	0.03816
<b>LIWC Cognitive</b>	
fraction of cognitive processes	0.01288
fraction of insight	-0.12494
fraction of causation	0.07551
fraction of discrepancy	0.05359
fraction of tentative	0.01638
fraction of certainty	0.15804
fraction of inhibition	0.04712
fraction of inclusive	-0.02781
fraction of exclusive	0.0368
<b>Readability</b>	
Flesch Kincaid grade	-0.0112

**Supplementary Table 3:** Fatigue model feature coefficients.

to human-written notes compared to the greedy decoding method, which has a higher readability and comes with a higher fraction of exclusive words in LIWC, e.g., “but” and “without”.

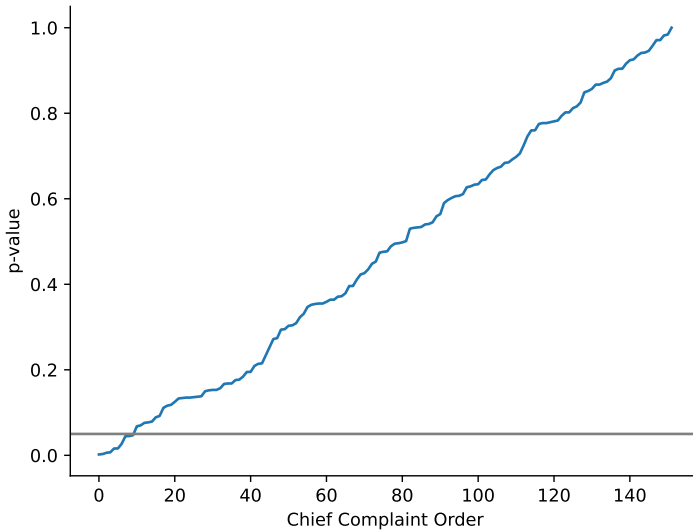
### 3.2 Generated Notes Have Higher Fatigue Scores

We generate the History of Present Illness (HPI) section of the notes with our fine-tuned GPT-2 model. For each note, we truncate it at the HPI section heading and generate the HPI section with greedy and sample decoding. We then obtain the fatigue score of the generated and original HPI sections with the fatigue predictor using the same set of features. Supplementary Table 7 shows that the predicted fatigue score is higher for the generated notes. We observe that generated notes have higher log perplexity scores for both greedy

	(1) Is Female	(2) Is White	(3) Age
Workload	-0.0055 (0.006)	-0.0023 (0.006)	-0.1562 (0.189)
Intercept	0.5694*** (0.029)	0.4724*** (0.03)	40.333*** (0.961)
<b>Controls</b>			
Time of Day	YES	YES	YES
Day of Week	YES	YES	YES
Week of Year	YES	YES	YES
Year	YES	YES	YES
Chief Complaint	YES	YES	YES
Physician	YES	YES	YES

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

**Supplementary Table 4:** Sanity checks of regressing patient demographics on the high- and low-workload indicator on all cohorts ( $n = 43,720$ ). The regression results show that the target label (workload indicator) of our predictive model is not correlated with patient demographics.  $p$ -values of workload coefficients are 0.343, 0.699, and 0.41 for column 1, column 2, and column 3 respectively based on two-sided tests.



**Supplementary Figure 1:** Sanity check for chief complaints. In only 10 out of 154 chief complaints, the workload coefficients have a significant effect, which represent 6.5% of all chief complaints in the entire dataset, close to 5% expected by chance at a significance level  $p=0.05$ .

Feature	Mean	Stderr
<b>Note characteristics</b>		
log perplexity	1.0170	0.0097
note length	1060.4227	12.6973
fraction of stopword	0.1927	0.0014
fraction of medicalword	0.3834	0.0012
<b>LIWC Pronoun</b>		
fraction of pronoun	0.0222	0.0003
fraction of i	0.0012	0.0001
fraction of we	0.0003	0.0000
fraction of you	0.0000	0.0000
fraction of shehe	0.0116	0.0002
fraction of they	0.0001	0.0000
fraction of ipron	0.0088	0.0001
<b>LIWC Affect</b>		
fraction of affect	0.0181	0.0002
fraction of posemo	0.0076	0.0001
fraction of negemo	0.0104	0.0002
fraction of anx	0.0015	0.0000
fraction of anger	0.0008	0.0000
fraction of sad	0.0033	0.0001
<b>LIWC Cognitive</b>		
fraction of cogmech	0.0858	0.0005
fraction of insight	0.0106	0.0001
fraction of cause	0.0074	0.0001
fraction of discrep	0.0044	0.0001
fraction of tentat	0.0120	0.0002
fraction of certain	0.0050	0.0001
fraction of inhib	0.0052	0.0001
fraction of incl	0.0331	0.0003
fraction of excl	0.0128	0.0002
<b>Readability</b>		
Flesch Kincaid score	10.0976	0.0816

**Supplementary Table 5:** Feature values on admission notes, charted within 6 hours of admission, in MIMI-III physician notes ( $N = 1,216$ ).

Feature	Original	Greedy	Sampling
<b>Note characteristics</b>			
log perplexity	1.614 (0.009)	0.598 (0.003)	1.334 (0.007)
note length	115.869 (1.506)	94.287 (1.014)	96.05 (1.087)
fraction of stopword	0.462 (0.002)	0.502 (0.002)	0.46 (0.002)
fraction of medical word	0.482 (0.002)	0.457 (0.002)	0.49 (0.002)
<b>LIWC Pronoun</b>			
fraction of pronoun	0.104 (0.001)	0.11 (0.001)	0.099 (0.001)
fraction of i	0.001 (0.0)	0.0 (0.0)	0.0 (0.0)
fraction of we	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
fraction of you	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
fraction of shehe	0.059 (0.001)	0.065 (0.001)	0.056 (0.001)
fraction of they	0.001 (0.0)	0.0 (0.0)	0.0 (0.0)
fraction of ipron	0.043 (0.001)	0.044 (0.001)	0.041 (0.001)
<b>LIWC Affect</b>			
fraction of affect	0.04 (0.001)	0.034 (0.001)	0.04 (0.001)
fraction of posemo	0.008 (0.0)	0.004 (0.0)	0.006 (0.0)
fraction of negemo	0.032 (0.0)	0.03 (0.001)	0.033 (0.001)
fraction of anx	0.005 (0.0)	0.002 (0.0)	0.003 (0.0)
fraction of anger	0.001 (0.0)	0.0 (0.0)	0.001 (0.0)
fraction of sad	0.003 (0.0)	0.006 (0.0)	0.005 (0.0)
<b>LIWC Cognitive</b>			
fraction of cogmech	0.127 (0.001)	0.12 (0.002)	0.126 (0.001)
fraction of insight	0.011 (0.0)	0.004 (0.0)	0.009 (0.0)
fraction of cause	0.005 (0.0)	0.001 (0.0)	0.004 (0.0)
fraction of discrep	0.003 (0.0)	0.002 (0.0)	0.002 (0.0)
fraction of tentat	0.031 (0.001)	0.045 (0.001)	0.034 (0.001)
fraction of certain	0.002 (0.0)	0.0 (0.0)	0.002 (0.0)
fraction of inhib	0.008 (0.0)	0.013 (0.001)	0.01 (0.0)
fraction of incl	0.055 (0.001)	0.043 (0.001)	0.053 (0.001)
fraction of excl	0.026 (0.0)	0.037 (0.001)	0.028 (0.001)
<b>Readability</b>			
Flesch Kincaid Score	8.5 (0.069)	7.548 (0.18)	8.345 (0.075)

**Supplementary Table 6:** Feature values of original HPI section and model generated HPI with different decoding methods ( $n = 1471$ ).

	Predicted Fatigue Deviation from the Mean	Log Perplexity	Fraction of Anger Words
<b>Decoding method</b>			
Greedy	189%	0.5983	0.03%
Sampled	8%	1.3335	0.12%
Original HPI	-66%	1.6141	0.09%

**Supplementary Table 7:** Fatigue score of generated notes with different decoding methods. Greedy decoding generates notes with higher fatigue scores than sampled decoding. Both types of generated notes have higher fatigue scores than the original HPI notes. Results are based on 1471 sampled notes from the hold-out set.

and sample decoding, and the fraction of anger words increases for sample decoding, which is the standard practice for recent large language models, e.g., GPT-4.<sup>1</sup>

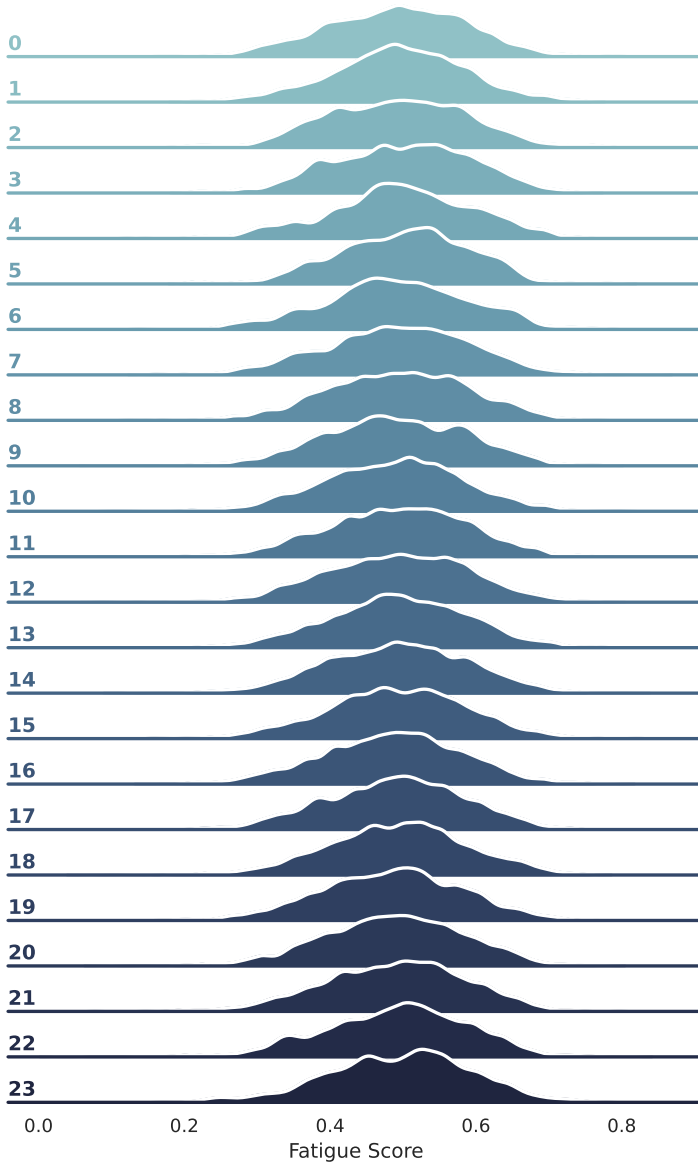
## 4 Feature importance

Supplementary Table 8 shows the full table of feature importance based on Pearson’s correlation with the high- vs. low-workload indicator, the yield of testing indicator, and the non-White indicator.

---

<sup>1</sup>The default top\_p value is 1 in the OpenAI API documentation (<https://platform.openai.com/docs/api-reference>). Results are robust with different top\_p.





**Supplementary Figure 2:** Fatigue score distribution over time of day on the hold-out set ( $N = 34,175$ ).

Feature	Pearson's Correlation	P-value
<b>Note characteristics</b>		
note length	-0.013861	3.75e-03
log perplexity	-0.092215	5.34e-83
fraction of stopwords	-0.008762	6.69e-02
fraction of medical words	0.005231	2.74e-01
<b>LIWC Pronoun</b>		
fraction of pronoun	-0.017585	2.35e-04
fraction of first person singular pronouns	-0.069399	7.85e-48
fraction of first person plural pronouns	0.017109	3.46e-04
fraction of second person pronouns	-0.000396	9.34e-01
fraction of third person singular pronouns	0.00593	2.15e-01
fraction of third person plural pronouns	0.00497	2.99e-01
fraction of impersonal pronouns	-0.046246	3.82e-22
<b>LIWC Affect</b>		
fraction of affective processes	-0.004541	3.42e-01
fraction of positive emotions	-0.003531	4.60e-01
fraction of negative emotions	-0.005992	2.10e-01
fraction of anxiety	-0.001283	7.89e-01
fraction of anger	0.024927	1.86e-07
fraction of sadness	0.003099	5.17e-01
<b>LIWC Cognitive</b>		
fraction of cognitive processes	0.025144	1.45e-07
fraction of insight	-0.087513	4.34e-75
fraction of causation	0.024334	3.60e-07
fraction of discrepancy	0.027211	1.26e-08
fraction of tentative	0.028669	2.02e-09
fraction of certainty	0.075771	1.07e-56
fraction of inhibition	0.037426	4.92e-15
fraction of inclusive	0.007204	1.32e-01
fraction of exclusive	0.040667	1.78e-17
<b>Readability</b>		
Flesch Kincaid grade	-0.04833	4.88e-24

**Supplementary Table 8:** Full table of feature importance based on Pearson's correlation with the high- vs. low-workload indicator ( $n = 43,730$ ).  $p$ -values come from two-sided tests. We use the name of the corresponding LIWC category to represent the name of a set of lexical words, so "fraction of inclusive" should be interpreted as "fraction of the inclusive category".