# Clinical Note Generation to Address Physician Burnout

Stanford CS224N Custom Project

**Jiying Zou**
Department of Statistics
Stanford University
jiyingz@stanford.edu

External Mentor: Diego Saldana (diego.saldana@roche.com)
TA Mentor: Akshay Smit (akshaysm@stanford.edu)

## Abstract

Physician burnout is one of the main contributors to the decreasing quality and personalized touch of clinical visits. Burnout in large part can be attributed to tedious and inefficient documentation processes of current EHR systems, as each visit can cost several hours of documentation afterwards. In this paper, we aim to build a language model to help generate clinical note contents, which could be deployed as part of an auto-complete system to increase efficiency of the documentation process. We believe that we are one of the first to deploy GPT-2 architecture for this objective. Results show that a small GPT-2 model finetuned on the MIMIC-III clinical note corpus can replicate note structure quite dependably and often fills in contents of reasonable length and semantic appropriateness. The same model struggles to handle medical abbreviations, special characters, and nuanced formatting, illustrating the importance of data quality pre-processing. Appending additional diagnosis history did not improve model performance. Our findings provide evidence of feasibility for using such models in text-prediction software under real-life clinical settings.

## 1 Introduction

While medical advancements grow at an ever increasing pace, the quality of healthcare has incongruously lagged further and further behind. A major factor contributing to the steady decline in high-quality clinical interactions is physician burnout – the phenomenon where the demanding and stressful occupational pace-of-life renders physicians feeling emotionally worn out, demotivated, and detached from their patients on a personal level. The current inconvenient and inefficient structure of electronic health records (EHRs) play a major role in physician burnout, as they necessitate excessive data entry and clinical note-taking [1]. The medical scribe industry has blossomed in response to meet tedious behind-the-scenes documentation needs. Medical scribes are non-professionals hired to help fill in the paperwork behind-the-scenes for clinical interactions and patient records. The issue lies not only in that scribes often lack important medical expertise to handle such a critical element of care, but also that the lower-quality documentation is vulnerable to leakages of external influences (for example financial or administrative interests) [2]. Improved technologies assisting clinical documentation could go a long way in helping reduce burnout and motivating higher-quality physician-patient interactions.

Within this space, automated clinical note generation is still a rather under-explored opportunity. A study showed that physicians can spend up to 2 hours on EHR documentation for every hour spent face-to-face with a patient [3]. The fact that significant portions of a doctor's day is spend documenting interactions naturally incentivizes shortcuts in the process, the most prominent being

recycling notes from one patient to another. As Deborah Nelson, MD, succinctly summarizes, such copy-paste-and-modify habits lead to clinical notes that "are confusing, increasingly lengthy, uninformative, disorganized, internally inconsistent, misleading, or lacking in credibility and may introduce and propagate errors that can place patients at risk" [4]. If it were possible to introduce some auto-complete of sorts for the EHR documentation process similar to that provided while drafting emails, it could help re-motivate custom-tailored notes while addressing efficiency concerns. Our work here aims to contribute to the effort of increasing automation in EHR documentation while not detracting too much from its personalized, professional touch. We find that there is potential to do so via text-generative models that could assist in the live EHR documentation process.

## 2 Related Work

Clinical notes have been utilized in a variety of NLP tasks. A vast amount of such work has been focused on learning meaningful representations for clinical notes to then predict patients' risks for certain adverse events. Some examples include predicting 30-day hospital readmission rates [5] and mortality rates [6]. Another common objective is to extract useful bits of information which can then be used to automatically generate parts of EHR entries. One study used this approach to suggest diagnoses codes and patient notes within the EHR system, but focused on patient-clinician dialogues rather than on written notes [7]. A third direction involves generating plausible clinical notes from real-life references to be used for dataset augmentation where downstream tasks suffer from a paucity of data [8], but the goal leans towards originality rather than staying relevant to the source. Overall, surprisingly little work has been done in the space of learning from a patient's medical history to help generate future clinical notes. Recent research by Peter J. Liu of Google Brain (amongst others) has shown that incorporating relevant information about a patient along with contextual token hints to predict the probability distribution of subsequent words is significantly helpful on both local and global metrics, but there is a trade-off as traditional Transformer models may experience difficulty handling resulting long input sequences [9]. In this project, we aim to extend this research by exploring the utility of patient diagnoses codes as a succinct proxy for extensive medical context information.

## 3 Objectives

GPT-2 is one of the state-of-the-art generative text models that uses a decoder-only architecture and is pre-trained on a giant corpus of web-scraped text. It is able to model the probability of the next word given previous ones, which is the traditional language modeling task of estimating $P(w_n|w_{n-1}, ..., w_1)$ [10]. Despite its popularity, to our knowledge we are one of the first to employ it for medical note generation purposes, with the end intent of producing clinically-plausible notes. Our objectives are two-fold:

1. To determine which GPT-2 model version(s) can be meaningfully adapted to the clinical note generation task

2. To explore whether or not augmenting the notes with a patient's past diagnoses improves generated note plausibility

## 4 Methods

This project is an instance of conditional language modeling. Whereas in traditional language modeling we only feed models previous words/tokens $w_1, ..., w_{n-1}$, in conditional language modeling we seek to improve the next-word modeling process by providing additional context $C$ in addition to text token hints $w_1, ..., w_{n-1}$ [9]:

$$P(w_1, ..., w_n|C) = \prod_{i=1}^{n} P(w_n|w_{n-1}, ..., w_1, C)$$

Within our framework, the context $C$ given is a patient's prior diagnoses in the forms of numeric codes or descriptive text about diagnoses. The following diagram outlines our approach, and details follow.

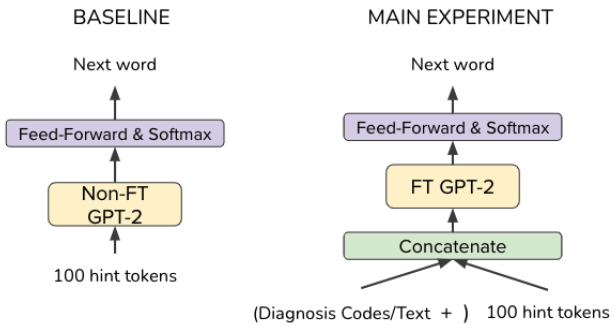


Figure 1: Sketch of baseline and experimental models. Main experiment diagram includes all three model variants – note text only, appended ICD-9 codes, and appended diagnosis text. "FT" is shorthand for "finetuned".

### 4.1 Baselines

Pre-trained language models contain a world of information on general semantic relationships from the corpus it was trained on, but may not perform as well in tasks that require more specific contextual adaptation. One example of such tasks is generating text with some uncommon formatting constraints and/or increased prevalence of specialized terminology, both of which is found in clinical notes. That said, we take our baseline models to be non-finetuned tiny, small, and medium versions of the GPT-2 model. The small and medium pre-trained versions of GPT-2 are publicly released by their authors at OpenAI, and the tiny version is an adaptation by Sam Shleifer, research engineer at Hugging Face. The model code and any released details can be found here.

### 4.2 Finetuning

We first finetune our tiny, small, and medium GPT-2 models on a corpus containing a mix of clinical note types (nursing, emergency, surgery, medicine, etc.), in hopes that the model will pick up on formatting and commonly used terms and abbreviations from these notes. All notes have similar, but not necessarily identical, formatting, and shorthand may vary by practice.

We also finetune our three models on augmented datasets (details to come in the section below). A common practice in the literature is to feed one finetuned model multiple augmented dataset versions, but we felt that additional finetuning was necessary as the model might need to learn associations of numeric diagnoses codes with note contents. Using each respective finetuned model, we feed in the corresponding test set corpus and generate notes for comparison and evaluation.

#### 4.2.1 Data Augmentation

Our clinical note corpus is augmented two ways. The first type of augmentation appends ICD-9 diagnosis codes (separated by spaces) for the same patient (from within a year prior) to the start of each note, delimited from the main text with the separating token <ICD>. The second type of augmentation is similar, but uses text diagnoses rather than their codes.

For illustration, a clinical note by itself may look like

```
Admission Date:  [**2118-12-7**]      Discharge Date:  [**2118-12-9**]...
History of Present Illness:  44 yo female with a h/o left frontal AVM...
```

The same note augmented with ICD-9 codes would be

```
73819 3485 34590<ICD> Admission Date:  [**2118-12-7**]      Discharge Date:
[**2118-12-9**]...  History of Present Illness:  44 yo female with a h/o
left frontal AVM...
```

and that note augmented using diagnosis in text format would be

```
Other specified acquired deformity of head Cerebral edema Epilepsy<ICD>
Admission Date:  [**2118-12-7**]      Discharge Date:  [**2118-12-9**]...
History of Present Illness:  44 yo female with a h/o left frontal AVM...
```

### 4.3 Text Generation

For each of our models, we feed in 100 hint tokens from each clinical note in the test set as input. For the models finetuned on augmented datasets, we correspondingly feed in the note concatenated with ICD-9 codes or diagnoses text in accordance with the style of the augmented finetune corpus. We chose 100 hint tokens because this was roughly the amount of information needed to inform the model of basic demographic information and purpose of visit, and possibly a couple starter words for the meat of the note.

Given past words $w_1, ..., w_{n-1}$, the GPT-2 model can produce a likelihood distribution over all words $x_i$ in its dictionary for $P(w_n = x_i)$. Many heuristics exist for how to pick the next word, but top-K and top-P sampling have shown the most promise for coherent text generation. In top-K sampling, as its name suggests, we randomly sample the next word from the $K$ words occurring next with highest probability [11]. The issue, though, with top-K sampling alone is that words in the tail end of the probability distribution may still be considered even though they are not likely choices; conversely, if the distribution is flatter, then many plausible words may not make the cutoff and are never considered. Top-P sampling offers a more flexible approach and addresses this issue by allowing consideration for any words (technically, tokens) whose probabilities are above some selected threshold. It is also possible to combine top-P and top-K sampling, for instance by sampling from the top $K$ words above some probability threshold $P$, which is the approach we take here.

## 5 Experiments

### 5.1 Data

Our clinical notes source from the MIMIC-III (Medical Information Mart for Intensive Care III) dataset, a freely-accessible data source containing medical records and demographic information from between 2001 and 2012 of critical care patients from Beth Israel Deaconess Medical Center [12]. We select a sample of the clinical notes from this database and use a 70%-20%-10% train-validation-test split, resulting in a datasets of sizes 5005, 1437, and 708, respectively.

### 5.2 Evaluation method

We evaluate our model performance using the local perplexity-per-token (PPL) metric and n-gram based ROUGE-1, ROUGE-2, and ROUGE-L scores. PPL is the standard language modeling evaluation metric that calculates the inverse probability of the corpus given a model, normalized by corpus length to mitigate for the fact that longer corpora yield intrinsically lower probabilities. It can also be expressed as the exponentiated average negative log probability of seeing the corpus $\hat{\mathbf{y}}$ up til each word $w_t$, given previous words $w_1, ..., w_{t-1}$.

$$perplexity = \Big(\prod_{t=1}^{T} \frac{1}{P(\hat{\mathbf{y}}_{w_t}|\hat{\mathbf{y}}_{w_1,...,x_{w-1}})}\Big)^{1/T} = exp\Big(\frac{1}{T}\sum_{t=1}^{T} -logP(\hat{\mathbf{y}}_{w_t}|\hat{\mathbf{y}}_{w_1,...,w_{t-1}})\Big)$$

ROUGE-1, ROUGE-2, and ROUGE-L measure the unigram, bigram, and longest common subsequence overlaps between the reference corpus (ground truth) and generated text. There are precision and recall-based variants for each measure [13]. Recall focuses on extent of content overlap with the reference text, while precision implicitly penalizes for verbosity in generated text. To be more lenient on succinctness, since note contents tend to vary a lot in length, we will report the F1-score, the harmonic mean of the two.

Our third criteria is human evaluation of generated text (1) coherence and (2) contextual plausibility since traditional quantitative metrics may be low even if the generated text is believable. We randomly sample 3 note indices from the test set and give the corresponding generated notes from each model to 6 human evaluators, all college-educated native speakers of English between the ages of 20 and 30 who are not medical professionals. They are asked to rank text coherence and contexual plausibility

on a scale of 1 to 10 each, with 10 being the best ranking. Rankings are collected using an anonymized survey with hashed model names and scrambled note order to avoid confounding. We report average rankings, since there seemed to be a rough concensus on the quality of most notes.

Any appended diagnosis information is removed during generated text evaluation.

### 5.3 Experimental details

All configurations and technical setups were consistent, but not optimal, per model for fair evaluation. We fine-tune pre-trained tiny, small, and medium GPT-2 models from the Hugging Face model repository for 5 epochs on our `TextDataset` objects using a block size of 128 tokens. We used mini-batch gradient descent for optimization, with a training batch size of 8 and evaluation batch size of 16. Evaluation happens every 400 steps. The learning rate is $5 \times 10^{-5}$, and we use the Adam optimizer with no weight decay. Tiny GPT-2 models take just under 2 hours to finetune, while GPT-2 small and medium models take nearly 15 and 40 hours per model, respectively.

Text generation is done using a mixture of top-P and top-K sampling, $p = 0.95$, $k = 50$. In consideration of time constraints, one note was generated per reference. Generating 708 notes corresponding to the test set takes roughly a day-and-a-half using GPT-2 medium, with time needed proportionately scaled down for the two smaller models.

### 5.4 Results

Here we present results from text generation evaluations. All evaluation was done on the test set. In the figures below, "FT" stands for a finetuned model, and the finetune corpus $\in$ {"Text" (clinical note text only), "ICD-9" (ICD-9 codes appended to note), "Diagnoses" (diagnosis text appended to note)}. For time constraint considerations, only a small GPT-2 (the most promising model that could run in time) was finetuned and evaluated on the dataset augmented with diagnosis text.

| | Evaluation Metric | | | | | |
|---|---|---|---|---|---|---|
| **GPT-2 Model** | PPL | ROUGE-1 | ROUGE-2 | ROUGE-L | Coherence | Plausibility |
| **Non-FT** | | | | | | |
| Tiny | 50242.2316 | 0.035 | 0.0278 | 0.0344 | 1.8 | 1.67 |
| Small | 33.5255 | 0.1711 | 0.0483 | 0.0957 | 5.73 | 3.13 |
| Medium | 22.8633 | 0.2439 | 0.0506 | 0.1348 | 1 | 1.07 |
| **Text FT** | | | | | | |
| Tiny | 809.6383 | 0.2442 | 0.0506 | 0.1348 | 1 | 1 |
| Small | 7.4756 | 0.2741 | 0.0855 | 0.1304 | 6.13 | 6.4 |
| Medium | 10.8276 | 0.2609 | 0.0877 | 0.1293 | 6.33 | 6.53 |
| **ICD-9 FT** | | | | | | |
| Tiny | 813.0863 | 0.2387 | 0.0505 | 0.1333 | 1 | 1 |
| Small | 13.0156 | 0.2477 | 0.0768 | 0.1217 | 4.87 | 5.07 |
| Medium | 47.8808 | 0.2322 | 0.0792 | 0.1221 | 4.2 | 3.93 |
| **Diagnoses FT** | | | | | | |
| Small | 7.8188 | 0.2501 | 0.0844 | 0.1238 | 4.87 | 4.93 |

Figure 2: Baseline and experimental evaluation metric results. Top 2 values are highlighted in each metric column (dark green = best value, light green = second best value).

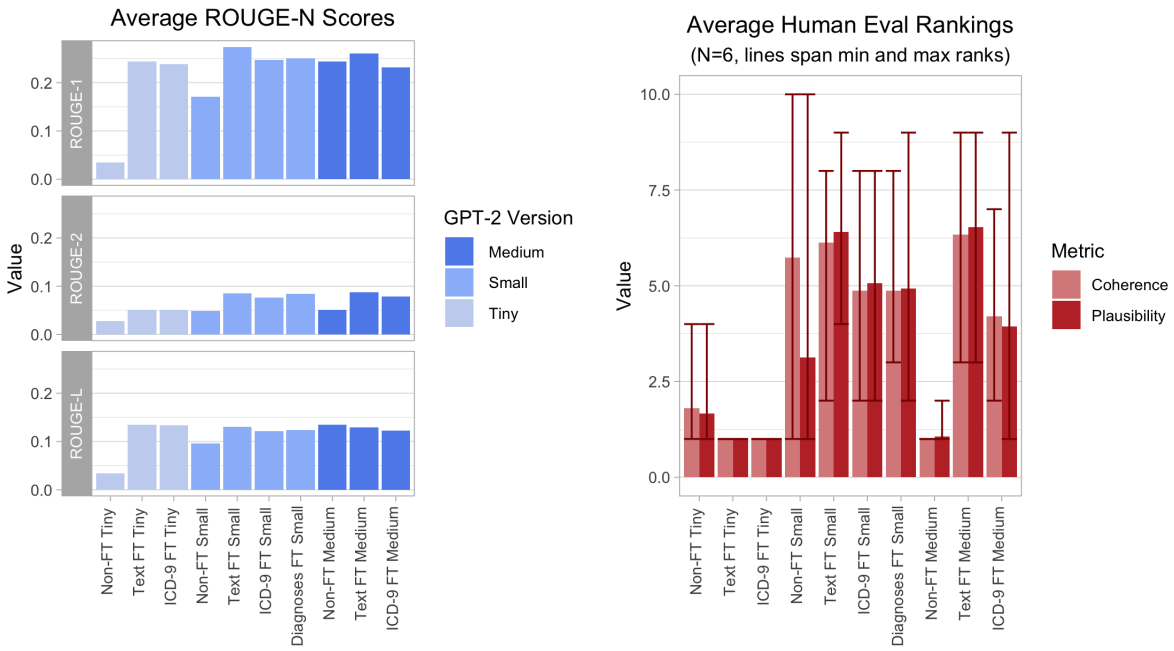


Figure 3: ROUGE and human evaluation metrics.

Finetuned models performed considerably better than non-finetuned models in most metrics, as expected. Small and medium GPT-2 models finetuned only on the clinical note text had highest performance across metrics overall and also produced the most coherent and plausible text. Feeding the model extra hints about diagnoses in either format, though, did not improve performance as expected, but rather slightly dented performance across the board. It is possible that finetuning separate models for each of the augmented datasets taught the model minimally useful associations while interfering with selection probabilities for other words, hence slightly upping perplexity.

It is also surprising that the medium GPT-2 variant performed either on par with (for ROUGE and human evaluation) or even worse (for perplexity) than the small variant. Medium GPT-2 has 345M parameters, which is over twice the amount that small GPT-2 does, and given our small training corpus of just over $\sim$5000 notes, the model could have overfit without learning many more useful relationships than small GPT-2 did.

## 6 Analysis & Discussion

Here we elaborate on generation quality by examining some examples of generated text. Without finetuning, the tiny and medium models produce gibberish, whereas the small GPT-2 model generates repeated patterns emulating note sections. There seems to be awareness of formatting, but attention weights were incorrectly focused (examples in the Appendix Exhibits A1-A3). There are strong hints of wording that sounded like drug labels or medical disclaimers, which were likely the closest things it could draw from in its pre-trained knowledge. After finetuning, the small and medium GPT-2 models were able to pick up general formatting in plausible ways. They could generate text chunks of reasonable length and semantic order per note section (e.g. for patient illness history this would mean first describing the patient's condition, then procedures and treatments, then medications), and could even catch more subtle custom within-text formats such as lists, which culminated in more believable output. There were occasionally chunks of text that focused terminology well in one or two disease areas (e.g. heart disease, chest pain, nausea), but there were also times when models still produced confusing semantics or were still repetitive, especially for within less commonly appearing subsections. This may indicate that further finetuning is necessary to pick up on these finer nuances, and/or that a larger training corpus is needed to provide more examples for rare subsections. On the topic of formatting, systematic list-like formattings seemed easier to pick up than ones that were more

free-form or involved symbols commonly used elsewhere. Examples of good and bad generations can be found in the Appendix Exhibits B1-B3.

One factor hindering better performance is likely our limited dataset and finetuning. The MIMIC-III database contains over 1 million notes, but within our time constraints we were only able to use a small fraction of that. This limits the amount of information our models can acquire and increases the danger of overfitting, especially on larger model variants. A second hypothesis is that while formatting may be one of our models' greatest successes, it may also be the factor holding back performance the most. There are lots of shorthand abbreviations (e.g. `CAD s/p MI`, `CABG [**2169**] (LIMA-LAD, SVG-RCA, SVG-D1-OM)`) and special characters (e.g. the anonymized portions that are in format `[**YYYY-MM-DD**]` or `[**HOSPITAL101**]` present in the reference data that we did not alter or remove, and are difficult even for humans to understand. Occassionally there will be hidden multimodality, where tables, checkboxes, or other specialized formats extending beyond natural language are hidden within the text, for example:

```
Rt      2+  3+  2+ 2+
Lt      2+  x   2+
```

In addition, the data itself has many spaces and newlines that are difficult to pre-process away without altering the fundamental structure of the notes, and may have contributed to model confusion. As a contrast, we conducted a similar side experiment to generate PubMed case reports using small GPT-2 with and without Medical Subject Headings (MeSH) appended to the corpus. We found that the text generated was much more believable, did not suffer from formatting issues, and that perplexity actually improved when the corpus was augmented with MeSH. This lends validity to our hypothesis that our data quality may be a big part of the problem since PubMed abstracts did not have the same stylistic issues that our note corpus had.

Furthermore, we hypothesize that appended diagnosis information was not helpful because numbers in ICD-9 codes are a form of encoded information in a way that does not overlap with how numbers are usually used or related to words, and so the model (or even the tokenizer) did not acquire pertinent deciphering abilities from the brief finetuning process. It may also be that some notes leaked the main medical complaint within the first 100 tokens, so re-iterating this knowledge did more harm than good. A third idea is that note type heterogeneity rendered the learning process less effective. Further research is needed to evaluate these claims.

## 7 Future Directions

Our aim from the beginning was to reduce physician burnout through an auto-completion system. While the finetuned small GPT-2 model may not yet be able to produce wholly clinically-plausible notes, it does show promise in suggesting the next couple words. For instance, it consistently gets subsection headings correctly, and is able to produce relevant text lengths and contents for many of them. Some samples are:

```
Allergies: Patient recorded as having No Known Allergies to Drugs

Major Surgical or Invasive Procedure: None

History of Present Illness: Mr. [**Known lastname 25191**] is a 68 year old male who
     presented to Dr.[**Name (NI) 2352**] office on [**5-6**] for hematoma
    evacuation, pancreatic transplant, and pancreatic mass resection after fall....

Past Medical History: Coronary Artery Disease
```

Under this application it is not necessary that the whole note be perfectly formed, but rather only tidbits at a time.

Before further development using this model, one should run experiments to finetune on a larger training sample for more epochs. It would also be interesting to see if additional appended demographic information or previous clinical notes will improve generated text quality.

A potential pitfall of such a system is that we currently do not have a way for the model to stop producing text or providing suggestions if it is uncertain about the next word. Medical terminology

is often complex and abbreviations may look similar, so the danger lies in producing seemingly legitimate text that actually may be inaccurate. Extra caution would be necessary to not accidentally pick the incorrect suggested diagnosis or treatment, which would contribute to misdiagnosis and incorrect future treatments. For deployment of such a system, words should not be suggested if they do not pass some threshold of confidence.

## 8 Conclusion

In this project, we sought to build a GPT-2-based language model to generate plausible clinical notes. We found that a finetuned small GPT-2 model was sufficient for replicating clinical note structure and general semantics. With our current work, we already observe snippets of on-topic, context-relevant medical summaries. The main limitations lie in effectively using abbreviations and more nuanced formatting, but finetuning with a larger training sample and better data pre-processing would likely improve generated text quality. Our research shows that GPT-2 holds promise in being deployed for a clinical note auto-completion software in the future, which holds implications in efforts to usher in AI/ML-based solutions to tackle the most pressing and widespread problems in healthcare today.

Code for this project is open-sourced at https://https://github.com/jiyingz/clinicalGPT-2.

## 9 Acknowledgments

Throughout this project, I used the Hugging Face library to work with Transformer models. We base our workflow and model finetuning code off of this tutorial by Phil Schmid on German recipe generation with GPT-2. I modified the `TextDataset` object from its source to circumvent system errors. Text generation code is modeled off of bits from a Hugging Face blog tutorial on generation, accessible here. Code for calculating perplexity is borrowed off of this guide from Hugging Face. ROUGE score calculation is done using the `rouge_score` package and relevant code is adapted from this guide. The side experiment concerning PubMed case report generation with MeSH was conducted entirely by Diego Saldana.

In addition to my mentors, I would also like to give thanks to Lauren Zhu for general advice during the beginning and end of this project, to the CS224N course staff for all their efforts this quarter, to all my friends for helping last-minute evaluate my model generations, to family and JJ for emotional support through hard times, to Microsoft Azure for project credits, and to my faithful virtual machine VirtyMcVirtFaceTheSecond for running countless hours of arduous code.

## References


[1] P. J. Kroth, N. Morioka-Douglas, S. Veres, S. Babbott, S. Poplau, F. Qeadan, C. Parshall, K. Corrigan, and M. Linzer. Association of Electronic Health Record Design and Use Factors With Clinician Stress and Burnout. *JAMA Netw Open*, 2(8):e199609, 08 2019.

[2] GA Gellert, R Ramirez, and SL Webster. The rise of the medical scribe industry: Implications for the advancement of electronic health records. *JAMA*, 313(13):1315–1316, 2015.

[3] B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, W. J. Tuan, C. A. Sinsky, and V. J. Gilchrist. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Ann Fam Med*, 15(5):419–426, 09 2017.

[4] Deborah D. Nelson. Copying and pasting patient treatment notes. *AMA Journal of Ethics*, 13(3):144–147, 2011.

[5] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342, 2019.

[6] Kexin Huang, Abhishek Singh, Sitong Chen, Edward T. Moseley, Chih-ying Deng, Naomi George, and Charlotta Lindvall. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. *CoRR*, abs/1912.11975, 2019.

[7] Faiza Khan Khattak, Serena Jeblee, Noah H. Crampton, M. Mamdani, and F. Rudzicz. Autoscribe: Extracting clinically pertinent information from patient-clinician dialogues. *Studies in health technology and informatics*, 264:1512–1513, 2019.

[8] Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, Sumithra Velupillai, and et al. Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, 3(1), 2020.

[9] Peter J. Liu. Learning to write notes in electronic health records. *CoRR*, abs/1808.02622, 2018.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[11] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. *CoRR*, abs/1805.04833, 2018.

[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[13] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

# A Appendix

Here we present some samples of generated text from various models. Note that the first 100 tokens were fed in rather than generated, which roughly corresponds to up until the "Service:" section of most notes.

**Exhibit A1.** Gibberish text generated by the non-finetuned tiny GPT-2 model.

```
Admission Date: [**2136-2-8**]  Discharge Date: [**2136-2-24**]
Date of Birth: [**2077-2-19**]   Sex: F

Service: Cardiothoracic Surgery

HISTORY OF PRESENT ILLNESS: This is a 59-year-old female
with a past medical history significant factors predators mutual soy perhaps448
    cleareraciousozyg predators Wheels Wheelsshows Late predators Late
    braverySexual rubbing clearer boils Late Late Tre equate Singaporepublic
    Boonepublic linedOutside Television Treozyg brutality brutality brutality
    Dreamsshows Tre653 Singapore Singapore mutual Redux grandchildrenMost
    membership predatorsozyg Dreams 236 membership soy representations workshops
    Bend soy skillet Wheels clearerGy soy boils Television Bend clearerPros Bend
    grandchildrenived653acious653Prosshows prayingMinipublicived courtyard Tre
    mutual Tre mutual linedshowsPros bravery mutual grandchildren Medic LateOutside
    ...
```

**Exhibit A2.** Repetitive text with some semblance of formatting generated by the non-finetuned small GPT-2 model.

```
Admission Date: [**2200-4-17**]          Discharge Date: [**2200-4-23**]

Date of Birth: [**2128-9-4**]          Sex:  M

Service: CARDIOTHORACIC

Allergies:
Iodine-Iodine

Amphetamine-Hydroxybenzoylmethane

Equalities:

Iodine: 2.2-5 times more common in adults than other drugs; 10.3 times more common
    among children 2 to 16 years of age

Equalities: 1.9 times more common for both sexes 3 times more common among teenagers
     and 18-21 year olds

Equalities:
```

**Exhibit A3.** Gibberish text generated by the non-finetuned medium GPT-2 model, which seemed to try too hard to utilize special characters and formatting.

```
Admission Date: [**2200-4-17**]          Discharge Date: [**2200-4-23**]

Date of Birth: [**2128-9-4**]         Sex:  M

Service: CARDIOTHORACIC

Allergies:
Iodine-Iodine Tablet2
, (1 on for.- patient
 to

**:, a1**- in. ],2 45- on* for.2)*

1 was. a. the [
.8.,
: Tablet1/: on/ and--
)]. and)* in was- and-,- to
**H,, was on** mg- Tablet

 at a

21
 ].
 to ::
**:**.**
```

**Exhibit B1.** A sample of a well-generated note by small GPT-2 finetuned on clinical note text only. Notice the subsections are in a plausible arrangement, are of reasonable length (albeit the line breaks are in odd places), and describes the patient's condition, treatment, and medications in order as real clinical notes tend to do. This particular note describes a woman with heart issues, and successfully used relevant terms like like CABG (coronary artery bypass grafting) and warfarin (a medication for preventing blod clots in those with heart function irregularities).

```
Admission Date: [**2136-2-8**]  Discharge Date: [**2136-2-24**]

Date of Birth: [**2077-2-19**]   Sex: F

Service: Cardiothoracic Surgery

HISTORY OF PRESENT ILLNESS: This is a 59-year-old female
with a past medical history significant for hypertension,
chronic obstructive pulmonary disease, atrial fibrillation, and
peripheral vascular disease who presents with
two days of chest pain radiating down the right arm. She reports that
the pain started the day of admission and has since spread
upthrough her arms to her back. Patient reports no blood loss or pain.
She had a cardiac catheterization the next morning which
showed 3 vessel coronary artery disease in one of three vessel disease. Patient
subsequently had one
echocardiogram in [**Month (only) 547**] of this year which showed EF of 55% and a
severe three vessel coronary artery disease in [**2140**] that has resolved
from the mitral valve and mitral valve stenosis
and an ejection fraction of 35% and a [**2-10**2135**] cardiac
inferior myocardial infarction. Patient presented in [**10/2128**], transferred from
     a car accident
secondary to a mechanical
ventricular bigeminy at the ED with no acute distress and
after a car accident in [**2125**] and was electively intubated. Patient was in
noted to have a non-Q wave myocardial infarction at the end of
[**Month (only) 404**]. Patient also developed
acute renal failure secondary to fluid overload on [**2125**] and was admitted to
the ICU. Patient presented to [**Hospital3 2560**]
and underwent an AVR on [**2127-2-10-21**] requiring 2.0 units of packed red blood
cells for
during hospitalization. Patient's ejection fraction of 55% to 65% which
increased to 80% in the
following
coronary angiogram on [**2127-2127-2-11**]. On [**2127-3-24**] patient ruled in for
with an
systolic congestive heart failure with
failure and
hypertension
exacerbation and
exacerbation from aortic valve and diastolic heart failure with cardiac
    catheterization
which showed two
and
which revealed no changes from prior cardiac
catheterization.
reportedly
s/negative per cardiac catheterization in [**2127-2-11-21**] with aortic
aneurysmal atrial
fibrillation, no
graft angiogram in [**2127-2-11**], treated with ECHO showed a normal left
saphenoid 50% stenosis. Patient was initially admitted
to the CCU but then called out of the CCU
after arrival at the [**Hospital3 2560**] Hospital.
.  She was noted to have an elevated troponin of 8 in [**1-28**], troponins negative
secondary to
hematuria in [**2-29**] and was found to have a new severe troponin of 12. Patient
    was
```

diagnosed with coronary artery disease and MI,
infiltrated on EKG. Cardiac catheterization at [**Hospital3 250**]
revealed 3 vessel disease. Patient was admitted to [**First Name4 (NamePattern1) **]
     [**Last Name (NamePattern1) **] at [**Hospital1 **]
[**Hospital1 **] and the CT angiogram showed 3 vessel coronary artery disease.
    Patient then had
CABG [**Hospital1 **]-Hemodialysis which was unsuccessful. She was then admitted to
the CCU for
management of cardiac
failure and was managed with Lopressor, aspirin, Lopressor and Neo.
CARDIAC: Patient was started on Lopressor 50 mg po
q d and was admitted with a LVEF of
30-35% on the morning of [**2127-2-27**] and was
transferred to the floor. She was placed on Lopressor and
also put on Neo. Cardiac catheterization on
[**2127-3-11**] revealed severe 2 vessel coronary artery
disease
and was in sinus rhythm

**Exhibit B2.** Some snippets of well-generated text formattings from the small-GPT model.

```
Review of systems:

(+) Per HPI
(-) Denies SOB, fevers/chills, cough, abdominal pain, headache, dyspnea. Denies N/V,
     paroxysmal nocturnal dyspnea.

(-) Denies sick contact (no known allergies). Denies CP. + sore throat. Denies cough
    , dyspnea.

...

PAST MEDICAL HISTORY:
1.  Hypertension, status post total knee replacement; status post craniotomy two
    weeks ago.
2.  Coronary artery disease, status post myocardial infarction [**2106**].
3.  Paroxysmal atrial fibrillation, status post
four catheterization in [**2108**]
at [**Hospital6 1126**] in [**3-11**] secondary to a right carotid stenosis status
    post
resection.
4.  Status post bilateral cataracts. Carotid artery stent in [**2110**] and
    [**2106**]. Status post left
femoral
arter bypass graft in [**2108**], status post
craniotomy with Dr. [**Last Name (STitle) **] [**Last Name (NamePattern4) **].
5.  History of diabetes mellitus.
6.  Peripheral vascular disease times three. Hypercholesterolemia.
7.  Noninsulin dependent diabetes mellitus
```

**Exhibit B3.** Some snippets of failed text formatting generations from the same small-GPT model.

```
.
.
On the floor, pt found to be afebrile. She was diaphoretic and complaining of nausea
    , +abd pain, +NS, abd weight stable. Pt c/w guaiac positive.
.  RIJ CVL draining clear of fluid. BS >40, +BS. Hct 32, c/o RUQ, HD negative. CVL
    draining flat.
.  RIJ CVL draining fluid.
.  Pt given 1U PRBC x1. Pt admitted to [**Hospital Unit Name 153**] for further care.
     Transferred to ICU w/o VSS and with plan
to get [**Location (un) **] for Hct check, then d/c'd. She had HD stable.
.

...


[**Hospital1 **] x 3 days as well (2 more days prior to HD]. Then [**12-29**] pt had
     a new pain in HD [**4-3**] pain at HD 5 and [**Hospital1 **]
[**12-30**] + N/V as well. His last BP is 150/130, [**10-11**] and [**10-4**],
    [**11-8**]. He took 50 mg [**10-11**]. He has been on aspirin which he reports
    as well. His last dose of
lasix [**10-9**] was 50 mg [**Hospital1 **].
```

**Exhibit C.** Some illogical text generated by the non-finetuned small GPT-2 model, illustrating safety concerns with generative models in general.

```
Prevention

If you have hepatitis B, avoid getting hepatitis B-related medications. These drugs
    may cause a side effect, if taken or if they have no other known side effects.
    These include, but are not limited to:

Hepatitis B: If you have hepatitis B-related complications, including low blood
    pressure or heart disease, you might develop liver problems.

Treatments

Treatments may work if the virus is removed from a person's body and given through a
     needle or other injection.
```