# Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization

**Bharath Chintagunta**      **Namit Katariya**      **Xavier Amatriain**      **Anitha Kannan**
jaic4@vt.edu      namit@curai.com  xavier@curai.com anitha@curai.com

## Abstract

In medical dialogue summarization, summaries must be coherent and must capture all the medically relevant information in the dialogue. However, learning effective models for summarization require large amounts of labeled data which is especially hard to obtain. We present an algorithm to create synthetic training data with an explicit focus on capturing medically relevant information. We utilize GPT-3 as the backbone of our algorithm and scale 210 human labeled examples to yield results comparable to using 6400 human labeled examples ($\sim$30x) leveraging low-shot learning and an ensemble method. In detailed experiments, we show that this approach produces high quality training data that can further be combined with human labeled data to get summaries that are strongly preferable to those produced by models trained on human data alone both in terms of medical accuracy and coherency.

## 1 Introduction

With increasing usage of telehealth platforms (Mann et al., 2020), large scale ecosystems of providers and patients have become apparent. This has exacerbated the need for comprehensive visit summaries of the medical dialogues by the attending practitioner in order to facilitate accurate hand-offs to other care providers or as a means of recording the interaction. However, having providers write summaries after each encounter is not only time consuming but also costly, limiting the scalability of telehealth platforms (Shanafelt et al., 2016)

In these settings, an automated summarizer that can assist the practitioners can be extremely valuable. However, an important challenge of end-to-end medical dialogue summarization is the lack of large scale annotated datasets. Annotation of medical dialogues is expensive and slow because they need to be curated by trained experts. This is further compounded by the fact that labeled data may not be publicly shared because of patient privacy concerns and HIPAA regulations.

Recent approaches to summarization (Qi et al., 2020; Zhang et al., 2019) use transfer learning where a pre-trained model (e.g. through self supervision of learning a language model) is fine tuned with a labeled dataset. However, fine-tuning still requires hundreds to thousands of labeled examples to obtain reasonable performance. Methods such as (Joshi et al., 2020) aim to partially overcome these issues through modeling strategies that directly learn important inductive biases from smaller amounts of data. In addition, (Joshi et al., 2020) also handled data sparsity by leveraging a key insight of sequential nature of information flow in a medical dialogue: global summary of the dialogue can be composed from local dialogue turns (snippets). This enables collecting training data for snippets as opposed to the full conversation - an insight, we use in our paper as well.

Recently, OpenAI developed GPT-3, a neural language model that is capable of natural language generation and completion of tasks like classification, question-answering, and summarization (Brown et al., 2020). The focus of that work is to enable task-agnostic and zero-shot or low-shot performance as opposed to a pre-trained model that needs to be fine-tuned separately on every downstream task. In this paper, we investigate the following question: *How can a low-shot learner such as GPT-3 be leveraged to scale training data for medical dialogue summarization models?* In answering this question within the context of GPT-3 as a black box proprietary API[1], we took into account multiple considerations:

- *Medical Correctness (Joshi et al., 2020)*: Medical summarization warrants high recall and therefore the summarizer should be good at (1) capturing all the medical information (med-

---

[1] https://beta.openai.com/

66

ications, symptoms, etc.) discussed in the dialogue and (2) discern all the affirmatives and negatives on medical conditions correctly (e.g. no allergies, having a cough for 2 days).

- *Privacy Concerns*: At inference time, an API call to external services such GPT-3 may not always be possible due to HIPAA and privacy concerns.

- *Practitioner in the loop*: The technique needs to be easily amenable to a feedback loop that allows for leveraging manually curated human labels. This feedback loop is extremely important because the diversity and the long tail of data distribution in medical dialogue means that there can be parts of the summary that need to be edited by practitioners for medical correctness and completeness. Note that these edits can be used as additional data for improving the underlying model.

Taking into account these considerations, this paper makes the following contributions (Figure 1 for a quick overview):

- We introduce a medically-aware GPT-3 data labeler, GPT-3-ENS , that combines medical knowledge and an ensemble of GPT-3 for the purpose of medical dialogue summarization.

- We introduce the idea of using GPT-3-ENS as a dataset generator to facilitate learning an in-house summarization model. Our experiments show that we can obtain the same performance as that of human labeled dataset with 30x smaller amount of human labeled data. With only 210 expert curated summaries and GPT-3 as a labeled data simulator, we can mimic the performance of a summarization model trained on 6400 expert curated summaries.

- By combining generated datasets from GPT-3-ENS with a human labeled dataset, we show that we can obtain better performance than models trained on either one of the data sources.

The rest of the paper is structured as follows: § 2 discusses related work, § 3 explores whether GPT-3 can be used directly for medical summarization, § 4 introduces our approach, § 5 and § 6 describe

our datasets and metrics respectively while § 7 illustrates our experiments. We end the paper with § 8 discussing our conclusions and future work.
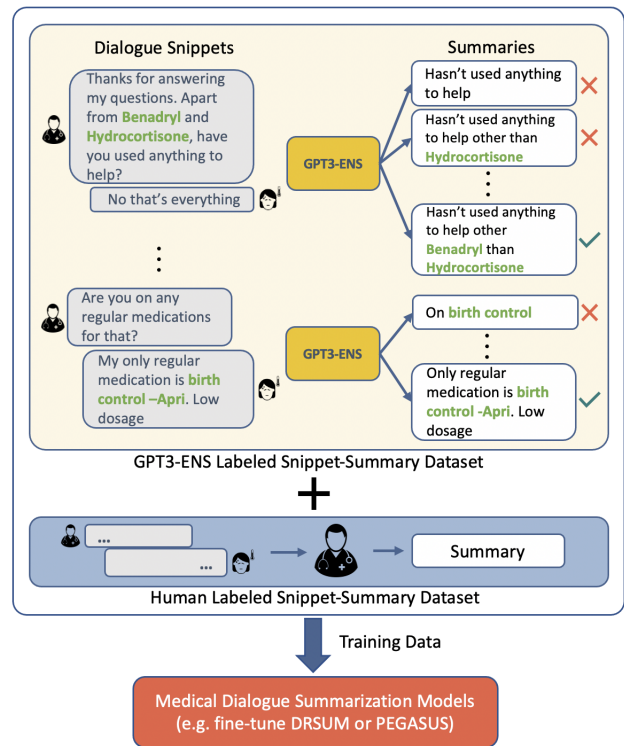


Figure 1: Overview of our proposed approach: we train models on a mix of GPT-3-ENS synthesized and human labeled data to get performance better than models trained on either of the sources

## 2 Related work

**Summarization** Emergence of sequence to sequence models and attention mechanisms (Sutskever et al., 2014) has led to rapid progress on extractive (Nallapati et al., 2017) , abstractive (Nallapati et al., 2016; Zhang et al., 2019) and hybrid models (See et al., 2017; Gu et al., 2016) for summarization. Much of the recent work has shown these models to generate near-human coherent summaries while retaining reasonable factual correctness.

**Dialogue summarization**: While most neural summarization has focused on news corpora, recent work has tried to tackle unique challenges associated with summarizing dialogues. (Goo and Chen, 2018) proposes using dialogue history encoders based on the type of dialogue section to inform the generation. (Liu et al., 2019a) propose using key points as a means of categorizing sections of dialogue.

**Medical dialogue summarization** Existing work

(Alsentzer and Kim, 2018; Zhang et al., 2018; Liu et al., 2019b; Krishna et al., 2020a,b; Joshi et al., 2020) in this space focuses on effective summarization by incorporating medical knowledge from a modeling perspective. Our work also focuses on incorporating medical knowledge from a data labeling perspective. We show how we leverage pretrained language models and low-shot learning (Brown et al., 2020) to collect labeled data for medical dialogue summarization. We also show how this data can improve performance over models that are trained solely on existing human labeled data.

## 3 Background: Can GPT-3 serve as a medical summarizer?

Ignoring the privacy concerns and practitioner-in-the-loop considerations, we first explore whether GPT-3 (Brown et al., 2020) is a good medical summarizer by itself.

GPT-3 takes as input a *priming context* to perform the task on a previously unseen example. Priming context refers to the text description of a task and a few demonstrations of the task being accomplished (in our case, that would be dialogue snippet summarization).

Table 1 column 2 provides examples of summaries generated by the GPT-3 model. We can clearly see that it misses a number of important pieces of information in the snippets – first, missing medical concepts making the summary unusable (Rows 1-2). Second, the model may not always get the affirmations correct (Row 3). Third, the summary may repeat redundant information from the doctor's queries (Row 4).

Based on these observations, we might *prematurely* conclude that GPT-3 can not be used for medical summarization task. However, our key observation in exploring GPT-3 is that it is sensitive to the priming context (also reported in (Liu et al., 2021)), as the model does not learn but just adheres to the examples given. As we show in 4, we exploit this variability in GPT-3 output via ensembling and infusion of medical knowledge so that it can be used as a part of an effective low-shot learning approach to medical summarization.

## 4 Infusing Medical Knowledge in GPT-3 for use as a Data Generator

We are interested in a model that uses only a small amount of human labeled data to learn an effec-

| Snippet | GPT-3 | GPT-3-ENS |
|---|---|---|
| **DR**: Thank you so much for sharing. Are you on any regular medications for that? **PT**: My only regular medication is birth control - Apri. Low dosage. | On birth control. | Only regular medication is birth control -**Apri**. **Low dosage**. |
| **DR**: You had mentioned the headache starting after stopping the medications. Have you had similar headache while stopping medications before? **PT**: Yes **PT**: Well that's a complicated question **PT**: I'm not really sure **DR**: Okay, no worries. Wanted to know a bit more since you had mentioned about them. That's all. | Yes. Stopping medications before. | Yes, has **headache** while stopping medications before. **Is not sure.** |
| **DR**: Do you have pain when you notice penile discharge? **PT**: no i'm not | Did not notice penile discharge. No pain. | **Doesn't have pain when noticing penile discharge.** |
| **DR**: I have a few questions to ask. How long have you been having this vaginal discomfort? **PT**: only a few days, since like friday or saturday maybe | I have a few questions to ask. How long has she been having vaginal discomfort? | Has been **having vaginal discomfort for only a few days, since friday or saturday.** |

Table 1: Input dialogue snippets along with summaries generated by GPT-3 in column 2 and our approach, GPT-3-ENS , in column 3.

tive medical dialogue summarizer. At the same time, we want such a model to be used in a practical practitioner-in-the-loop setting where medical correctness and patient privacy are of paramount importance.

In order to achieve these goals, we propose a two-pronged approach

1. Introduce **GPT-3-ENS** where we infuse medical knowledge into GPT-3 and use it within an inner loop to make it effective at medical summarization.

2. Leverage GPT-3-ENS as a data generator to obtain a large training set [2] to train an in-house medical dialogue summarization model. Such an in-house model can be used at inference

---

[2]Unlike data at inference time, training data is fixed and can be ensured to be privacy protected

68

time without the practical constraints related to protecting patient privacy that would require full de-identification to be applied in any conversation, if we were to access the GPT-3 service. It also lends itself well to the practioner-in-the-loop setting.

## 4.1 GPT-3-ENS : Medically-aware ensemble of GPT-3

As discussed in 3, GPT-3 is quite sensitive to the priming context. While one approach may be to provide GPT-3 with the most informative context for a task, this itself is a daunting task and can potentially be tackled if we had a large number of labeled examples (which is the exact problem we want to tackle with GPT-3).

Drawing on the learning from vast literature in ensembling techniques *c.f.* (Bishop et al., 1995), our first key insight is that if we can generate multiple summaries from GPT-3 using a variety of priming contexts, then we should be able to ensemble these outputs to identify the summary that is ideal for the dialogue. This insight leads to a question on how to ensemble multiple text summaries. The answer to this question relies on the core requirement for medical summarization: we care about the coverage of medical concepts mentioned and therefore the best ensembling function is the one that returns the summary with the most medical information in the dialog input.

In Algorithm 1 we provide our approach to the medically aware GPT-3 ensemble **GPT-3-ENS** . We assume access to a small set of labeled examples $\mathcal{L}$. For each input dialog snippet, $D$, we get $K$ summaries, by invoking GPT-3 each time with $N$ examples sampled randomly without replacement from $\mathcal{L}$. We also assume access to a medical entity extractor that can discern the medical concepts from both the dialogue snippet and the summary. The algorithm returns the best summary that has the highest recall in terms of capturing the medical concepts in the dialogue. For this purpose, we use an in-house medical concept extractor MEDICALENTITYRECOGNIZER that can identify medical concepts from a given piece of text. This extractor has access to the universe of medical concepts based on Unified Medical Knowledge Systems [3], which includes patient symptoms, disorders, laboratory tests and medications. Note that any medical entity recognizer (*cf.* (Fu et al., 2019) and references

---

[3] https://www.nlm.nih.gov/research/umls/index.html

therein) that has coverage for all these types of medical concepts found in medical conversations can be used.

---

**Algorithm 1** Medically aware GPT-3 ensemble summarizer (**GPT-3-ENS** )

---

**Require:** dialogue snippet $T$, ensembling trials $K$, universe $\mathcal{L}$ of labeled examples, medical entity extractor $MedicalEntityRecognizer$, GPT3
1: $C^* \leftarrow MedicalEntityRecognizer(T)$
2: **for** $i \leftarrow 1, \cdots, K$ **do**
3:     $S \leftarrow$ sample $N$ examples from $\mathcal{L}$
4:     summary$_i$ $\leftarrow$ GPT3$(S, T)$
5:     $C_i \leftarrow MedicalEntityRecognizer($ summary$_i)$
6: **end for**
7: $best \leftarrow \arg\max_i \frac{|C_i \cap C^*|}{|C^*|}$
8: **return** summary$_{best}$

---

Reconsider Table 1 for qualitative comparison between GPT-3 and GPT-3-ENS . We can see that summaries obtained using GPT-3-ENS capture the medical concepts comprehensively (shown in bold) and also have better grammatical structure. We also quantitatively validate the summaries on a small data set distinct from what is used for priming(see § 6.2 for guidelines). In Figure 2, based on doctor evaluation, we can see that GPT-3-ENS is significantly better at summarization than GPT-3 .
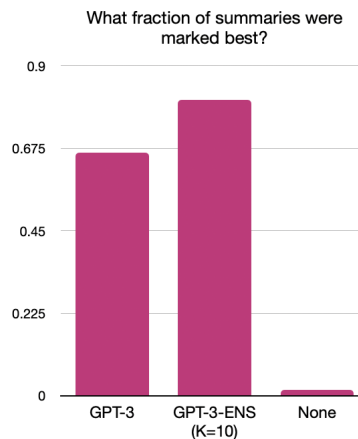


Figure 2: Doctor evaluation of which among GPT-3 and GPT-3-ENS summaries they considered "best" showing that GPT-3-ENS is a better approach for labeling

## 4.2 GPT-3-ENS as a data labeler

We use GPT-3-ENS described in 4.1 as our labeled data generator. In particular, we use our approach

to collect a large amount of labeled examples that serve as inputs to training an off-the-shelf summarization model. This resolves the concern of using GPT-3 in a real world application where the patient's conversation (in its raw form) needs to be exchanged with an external third party such as OpenAI/GPT-3 which may not have design/privacy regulations around HIPAA. In our approach, however, with the help of experts, it is easy to ensure that the dialogues that will used for priming as well as in the training set are chosen following privacy protocols.

## 5 Datasets

We collected a random subset of medical conversation dialogues from our chat-based telemedicine platform. Often medical conversation follows a linear ordering of medical history gathering (understanding patient symptoms) that enables creating the summary of the dialog by stitching together summaries of the snippets in chronological order (Joshi et al., 2020). Therefore, we split each dialogue into a series of local dialogue snippets using a simple heuristic: the turns between two subsequent questions by a physician corresponds to a snippet. The length of these snippets ranged anywhere from two turns (a physician question and patient response) to ten turns.

We had medical doctors[4] summarize these snippets. The doctors were asked to summarize the sections as they would for a typical clinical note by including all of the relevant history taking information. If a local snippet did not contain any history taking information it was excluded from annotations. For example in the beginning or end of conversations there may be turns that are purely greetings and not part of the patient history taking process. Further some snippets maybe purely educational in nature and are excluded as well. We eventually obtained a total of 6900 labeled snippet-summary pairs.

**Human labeled dataset train/test split:** From the 6900 labeled snippet-summary pairs (denoted as $H_{6900}$), we generated a randomly sampled test set $T = 500$ that we use in all our evaluations.

The dataset $H_{6900} - T$ is used to generate the priming dataset for GPT-3 related models as well as the datasets we use to train our summarization

models.

**GPT-3-ENS dataset**: Let $\mathbf{GCF}_p^k$ be the dataset of size $p$ generated using GPT-3-ENS with $k$ ensembling trials. To generate dataset $GCF^{K=k}$, we require $\{H_n\}_{i=1}^k$ datasets (note the independence on $p$), and thus $n \times k$ labeled examples for priming. These $n \times k$ examples are randomly sampled from the universe of human labeled examples $H_{6900} - T$. In our experiments, we sample without replacement so that no examples are reused across the $k$ tries. To allow comparison between our experiments with different $K$ values, we use the same seed for random sampling.

## 6 Evaluation Metrics

Multiple studies have shown that automated metrics in NLP do not always correlate well to human judgments as they may not fully capture coherent sentence structure and semantics (Stephen Roller, 2020; Kryściński et al., 2019). Since medical dialogue summarization would be used to assist health care, it is important for doctors to evaluate the quality of the output.

### 6.1 Automated metrics

While we measure model performance on standard metrics of ROUGE (Lin, 2004)[5], we also measure a model's effectiveness in capturing the medical concepts that are of importance, and their negations (Joshi et al., 2020)

**Medical Concept Coverage**: The concept coverage set of metrics captures the coverage of medical terms in the model's output summary with respect to the ground truth. In particular, let $\mathcal{C}$ be the set of medical concepts in the reference summary and $\hat{\mathcal{C}}$ be the set of concepts in the summary output by the model. Then, Concept recall $= \frac{\sum_{n=1}^{N} |\hat{\mathcal{C}}^{(n)} \cap \mathcal{C}^{(n)}|}{\sum_{n=1}^{N} |\mathcal{C}^{(n)}|}$ and Concept precision $= \frac{\sum_{n=1}^{N} |\hat{\mathcal{C}}^{(n)} \cap \mathcal{C}^{(n)}|}{\sum_{n=1}^{N} |\hat{\mathcal{C}}^{(n)}|}$.

We use these to compute a Concept F1[6] We use an in-house medical entity extractor to extract medical concepts in the summary. Medical concepts in the decoded summary that weren't present in the original conversation would be false positives and vice versa for false negatives.

---

**Negation Correctness**: To measure the effectiveness of the model to identify the negated status of medical concepts, we use Negex (Harkema et al., 2009) to determine negated concepts. Of the concepts present in the decoded summary, we evaluate precision and recall on whether the decoded negations were accurate for the decoded concepts and compute a negation F1[6].

## 6.2 Doctor Evaluation

We also had doctors, who serve patients on our telehealth platform, evaluate the summaries produced by the models. Given the local dialogue snippets and the generated summary, we asked them to evaluate the extent to which the summary captured factually correct and medically relevant information from the snippet. Depending on what percentage of the concepts were correctly mentioned in the decoded summary of the provided snippet, the doctors graded the summaries with *All* (100%), *Most* (at least 75%), *Some* (at least 1 fact but less than 75%), *None* (0%) labels.

We also formulated a comparison task where given summaries generated by different models and the associated dialogue, they were asked which summary was the "best" from a usability perspective. Usability was defined as whether the summary could stand in as a replacement for reading the dialogue snippet i.e. whether it captures the correct concepts from the snippet and whether the negations are accurate. The doctors had the ability to use "all" and "none" in this task depending on if all models being compared captured a good summary or if none of them did.

To avoid bias, the doctors do not know the model that produced the summary in both the experiments. In the comparison task, the summaries were provided in randomized order so that there is no bias in the order of presentation of the summaries.

## 7 Experiments and Results

**Additional models considered**: To evaluate the efficacy of **GPT-3-ENS** as a source of labeled data generator, we considered models with distinct objective functions for abstractive and hybrid (abstractive/extractive) summarization. We used **PEGASUS** (Zhang et al., 2019) for abstractive summarization and Dr. Summarize which we denote as **DRSUM** (Joshi et al., 2020) for extractive summarization. For **DRSUM** , we also use their best performing variant (referred as

2M-PGEN in (Joshi et al., 2020)) which penalizes generator loss and favors extractive copying.

**Implementation Details**: We used GPT-3 via the API released by OpenAI[7]. Maximum response length was set to 128 tokens, temperature to 0.6 and presence and frequency penalties both set to 0. For GPT-3-ENS , we use $K = 10$ ensembling trials for all our experiments, unless otherwise specified. We observed that $N = 21$ was the maximum number of examples we could prime GPT-3 with given the maximum context window length of 2048 tokens for the API. We therefore fix the size of our priming dataset to be 21 in all experiments which invoke GPT-3. Hence we set $L$ to be a random subset of 210 examples from $H_{6900} - T$.

We followed parameter settings for DRSUM from (Joshi et al., 2020) for pretraining on the CNN-Dailymail dataset. We then fine-tuned on our summarization task dataset with a batch size of 16, source_max_tokens = 400, response_max_tokens = 200 and max_grad_norm clipped at 2.0, for two epochs with a learning rate of 0.15 using Adagrad optimizer.

We used the PEGASUS implementation that is pretrained on CNN-Dailymail[8] provided by (Wolf et al., 2020). We fine-tuned it on our summarization task dataset with an effective batch size of 256, source_max_tokens = 512, response_max_tokens = 128 for two epochs using Adafactor[9] optimizer at the default settings in Hugging Face. For both PEGASUS and DRSUM , we used a beam size of four for decoding.

## 7.1 Training summarization models using data labeled by GPT-3-ENS

We compare PEGASUS and DRSUM trained on human labeled data $H_{6400}$ and GPT-3-ENS synthesized data $GCF_{6400}^{K=10}$. Note that synthesizing $GCF_{6400}^{K=10}$ needed all of $21 \cdot 10 = 210$ human labeled examples, where 21, as a reminder, is the maximum number of inputs that can be used for priming.

Table 2 compares quantitative performance of PEGASUS and DRSUM trained on these two datasets. The main observation is that with only

---

| Models | Train Data Source | Metrics | | |
|---|---|---|---|---|
| | | Negation F1 | Concept F1 | ROUGE-L F1 |
| PEGASUS | $H_{6400}$ | 21.09 | 35.96 | 55.59 |
| | $GCF_{6400}^{k=10}$ | **28.89** | 40.02 | 53.43 |
| | $GCF_{12800}^{k=10}$ | 26.70 | 40.21 | 56.66 |
| | $GCF_{25600}^{k=10}$ | 28.61 | **40.58** | **58.44** |
| DRSUM | $H_{6400}$ | **26.75** | **39.95** | **52.70** |
| | $GCF_{6400}^{k=10}$ | 24.29 | 37.55 | 48.47 |
| | $GCF_{12800}^{k=10}$ | 26.66 | 38.49 | 49.18 |
| | $GCF_{25600}^{k=10}$ | 26.08 | 39.47 | 50.85 |

Table 2: Automated evaluation of summarization models trained with different data labeling methodologies. Note that the amount of human labeled data is still pretty low (210), compared to 6400 when we do not use our approach.

210 human labeled examples, our approach GPT-3-ENS is able to generate a large amount of training data for both pre-trained summarization models, PEGASUS and DRSUM , in such a manner that they yield comparable (or better perfomance) than if they had been trained with only 6400($\sim$30x) human labeled examples.

For PEGASUS , the summarization performance improves drastically compared to model fine-tuned using only the human labeled data. We hypothesize that data generated from GPT-3-ENS can serve as quality training data for abstractive models such as PEGASUS but not so much for hybrid models such as DRSUM due to GPT-3 being a generative language model. The summaries written by our human doctors have writing structure similar to that of a hybrid summarization model such as DR-SUM that is more extractive in nature. This can explain why DRSUM did not show performance gain when using generated data from GPT-3-ENS . The key, however, is that it still did perform *on par*.

In the same Table 2, we also present the results with increased amounts of data (12800 and 25600) from GPT-3-ENS . There is little or no further improvement in the automated metrics of concept and negation F1. However, ROUGE-L F1 improves reflecting the improvements in coherency of the summaries. We leave this area as future work to explore.

### 7.2 Effect of combining human labeled data with data labeled by GPT-3-ENS

Since GPT-3 relies on limited local priming context ($N = 21$) it may not be agile in providing robust summaries for a multitude of variations in

snippets, focusing on the exploitation part of the exploration-exploitation trade-off. We hypothesize that best summaries then will be synthesized by a model trained on a dataset with human and GPT-3-ENS labeled examples. To evaluate this, we introduced a mixing parameter $\alpha$, the ratio of GPT-3-ENS labeled examples to human labeled examples. For instance, with 6400 human labeled examples, $\alpha = 0.5$ implies the dataset contains 6400 human labeled examples along with $0.5 * 6400 = 3200$ GPT-3-ENS generated examples. We experiment with $\alpha = 0.5, 1, 2, 3$.

From Table 4, we observe that for both PEGASUS and DRSUM , mixture of human labeled and GPT-3-ENS data consistently improves almost all automated metrics for all $\alpha$ values[10] The lift in metrics is lower for DRSUM , again illustrating the idea we highlighted in § 7.1 of GPT-3-ENS data being more amenable to abstractive models such as PEGASUS than for hybrid or extractive-biased models such as DRSUM . Table 3 provides qualitative comparison between summaries generated by each of these models.

For simplicity, we chose the smallest GPT-3-ENS mix i.e. $\alpha = 0.5$ for human evaluation where we ask doctors to evaluate summaries from model trained on human, GPT-3-ENS and human+GPT-3-ENS data. Figure 3 and Figure 4 show that doctors prefer summaries from the model trained on the mixture data over those produced by models trained on human or GPT-3-ENS data alone, in terms of amount of medical information captured as well as the overall quality of the summary. Furthermore, Figure 3(b) also shows that for PEGASUS , doctors prefer the summaries from a model trained on $GCF_{6400}^{K=10}$ (which needed *only 210 human labeled examples*) over those produced by a model trained on 6400 human labeled examples.

## 8 Conclusion

We introduced a medically-aware GPT-3 data labeler, GPT-3-ENS , for the task of medical conversation summarization. At the heart of the approach is a medically aware ensembling criterion that ensembles multiple summaries for an input from a powerful low-shot learner such as GPT-3. We showed that this approach can generate quality

---

[10]Note here that the claim is not that increasing $\alpha$ improves metrics but that mixing GPT-3-ENS and human labeled data improves metrics over models trained only using human data. We leave it as a future work on how to trade-off between human and GPT-3-ENS labeled data.

| Snippet | Model trained on $H_{6400}$ | Model trained on $GCF_{6400}^{K=10}$ | Model trained on $H_{6400}$+$GCF_{3200}^{K=10}$ |
|---|---|---|---|
| **DR**: Have you ever been tested for any underlying health conditions such as diabetes, hypothyroidism or polycystic ovarian syndrome? **PT**: No **PT**: I have been told I have prediabetes | Has not been tested for any underlying health conditions. | Hasn't tested for any underlying health conditions such as diabetes, hypothyroidism or polycystic ovarian syndrome | Has not been tested for any underlying health conditions. Has been told has prediabetes. |
| **DR**: DR: Do you have pus appearing discharge from the site? **PT**: Yes. If the bubbles pop it leaks out a watery substance | Has pus appearing from the site. | Pus appearing from the site | Pus discharge from the site. If bubbles pop it leaks out a substance. |

Table 3: Input conversation snippets along with summaries generated by models trained on different data

| Models | Train Data Source | Metrics | | |
|---|---|---|---|---|
| | | Negation F1 | Concept F1 | ROUGE-L F1 |
| PEGASUS | $H_{6400}$ | 21.09 | 35.96 | 55.59 |
| $\alpha = 0.5$ | $H_{6400} + GCF_{3200}^{K=10}$ | 30.14 | 43.49 | **62.45** |
| $\alpha = 1$ | $H_{6400} + GCF_{6400}^{K=10}$ | 30.70 | 43.73 | 60.63 |
| $\alpha = 2$ | $H_{6400} + GCF_{12800}^{K=10}$ | 29.43 | 41.02 | 59.85 |
| $\alpha = 3$ | $H_{6400} + GCF_{25600}^{K=10}$ | **31.93** | **44.68** | 61.05 |
| DRSUM | $H_{6400}$ | 26.75 | 39.95 | 52.70 |
| $\alpha = 0.5$ | $H_{6400} + GCF_{3200}^{K=10}$ | **27.51** | 40.46 | **53.39** |
| $\alpha = 1$ | $H_{6400} + GCF_{6400}^{K=10}$ | 27.18 | 40.36 | 51.00 |
| $\alpha = 2$ | $H_{6400} + GCF_{12800}^{K=10}$ | 27.19 | **40.68** | 53.07 |
| $\alpha = 3$ | $H_{6400} + GCF_{25600}^{K=10}$ | 26.33 | 39.89 | 52.29 |

Table 4: Combining human labeled datasets with datasets generated using our proposed approach

Figure 4: Doctor evaluation of amount of medical information covered by summaries provided by DR-SUM models and which ones they considered "best"
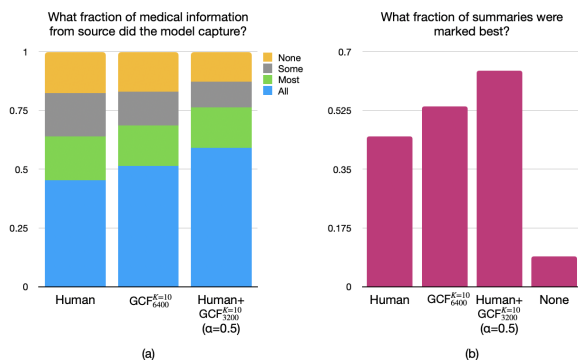
Figure 3: Doctor evaluation of amount of medical information covered by summaries provided by PEGA-SUS models and which ones they considered "best"
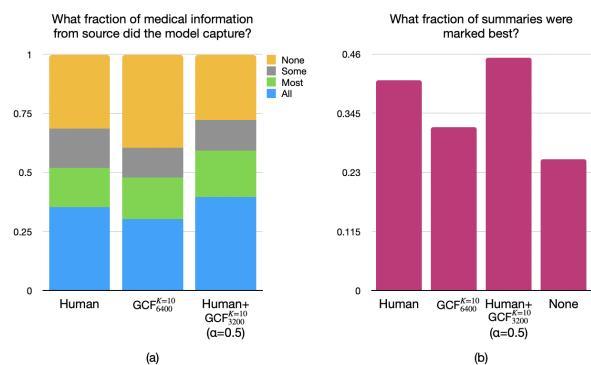
training data for medical dialogue summarization models while ensuring medical correctness. We show that using a very small number of human labeled examples, 210, we are able to produce more medically correct and better quality summaries than using roughly thirty times as many human labeled examples for two different summarization models. In this work we used a simple ensembling technique that dialogue summaries should retain all the medical information discussed in the dialogue. Future work could be to improve our ensembling function to take into account other medical priors such as affirmations and importance/relevance of the information in the dialog.

## References

Emily Alsentzer and Anne Kim. 2018. Extractive summarization of EHR discharge notes. *CoRR*, abs/1810.12085.

Christopher M Bishop et al. 1995. *Neural networks for pattern recognition*. Oxford university press.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sunyang Fu, David Chen, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Yanshan Wang, Liwei Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. 2019. Development of clinical concept extraction applications: A methodology review. *CoRR*, abs/1910.11377.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. *CoRR*, abs/1809.05715.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839 – 851. Biomedical Natural Language Processing.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Kundan Krishna, Sopan Khosla, Jeffrey P. Bigham, and Zachary C. Lipton. 2020a. Generating soap notes from doctor-patient conversations.

Kundan Krishna, Amy Pavel, Benjamin Schloss, Jeffrey P Bigham, and Zachary C Lipton. 2020b. Extracting structured data from physician-patient conversations by predicting noteworthy utterances. In *Explainable AI in Healthcare and Medicine*, pages 155–169. Springer.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization.

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

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019a. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page

1957–1965, New York, NY, USA. Association for Computing Machinery.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3?

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019b. Topic-aware pointer-generator networks for summarizing spoken conversations.

Devin M Mann, Ji Chen, Rumi Chunara, Paul A Testa, and Oded Nov. 2020. COVID-19 transforms health care through telemedicine: evidence from the field. *Journal of the American Medical Informatics Association*. Ocaa072.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training.

Abigail See, Peter Liu, and Christopher Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Tait D. Shanafelt, Lotte N.Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P. West. 2016. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic Proceedings*, 91:836–848.

Naman Goyal Da Ju Mary Williamson Yinhan Liu Jing Xu Myle Ott Kurt Shuster Eric M. Smith Y-Lan Boureau Jason Weston Stephen Roller, Emily Dinan. 2020. Recipes for building an open-domain chatbot.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. *CoRR*, abs/1809.04698.

| Snippet | Summary | Prompt |
|---|---|---|
| **PT:** Today spit out a bit of mucus and noticed a bit of blood.<br>**DR:** Okay, how long have you been on these medications?<br>**PT:** About 2 years | Has been on these medications for about 2 years | Today spit out a bit of mucus and noticed a bit of blood.[STOP] Okay, how long have you been on these medications?[SEP]About 2 years[SUMMARIZED]Has been on these medications for about 2 years.[STOP] |
| **DR:** Is the bleeding from the anal opening and not the vagina? Has something similar happened before?<br>**PT:** yes from the anal opening | The bleeding is from the anal opening. | Is the bleeding from the anal opening and not the vagina? Has something similar happened before?[SEP]yes from the anal opening[SUMMARIZED]The bleeding is from the anal opening.[STOP] |

Table 5: Prompt for GPT-3 given two examples

## A  GPT-3 Prompt

We utilize a fairly simple prompt to have GPT-3 generate summaries. Each example (snippet_text, summary_text) is concatenated to the empty string with the following transformation: "{snippet_text}[SUMMARY]{summary_text}[STOP]" to form the prompt. We seperate the conversational turns in snippet_text with the "[SEP]" token. Table 5 shows a prompt that would be generated and used to prime GPT-3 given two examples. As mentioned in § 7 in our experiments we use 21 examples to generate a prompt