CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models

Steven Y. Feng[1]*, **Vivek Khetan**[2], **Bogdan Sacaleanu**[2], **Anatole Gershman**[3], **Eduard Hovy**[3]
[1]Stanford University, [2]Accenture Labs, SF, [3]Carnegie Mellon University
syfeng@stanford.edu
{vivek.a.khetan,bogdan.e.sacaleanu}@accenture.com
{anatoleg,hovy}@cs.cmu.edu

Abstract

We motivate and introduce **CHARD**: Clinical Health-Aware Reasoning across Dimensions, to investigate the capability of text generation models to act as implicit clinical knowledge bases and generate free-flow textual explanations about various health-related conditions across several dimensions. We collect and present an associated dataset, CHARDAT, consisting of explanations about 52 health conditions across three clinical dimensions. We conduct extensive experiments using BART and T5 along with data augmentation, and perform automatic, human, and qualitative analyses. We show that while our models can perform decently, **CHARD** is very challenging with strong potential for further exploration.

1 Introduction

Pretrained language models (PLM) have seen increasing popularity for NLP tasks and applications, including text generation. Researchers have become interested in the extent to which PLMs can: 1) act as knowledge bases, 2) reason like humans.

Rather than using external databases, exposure to large amounts of data during training combined with their large number of parameters, has given PLMs the ability to store knowledge that can be extracted through effective probing strategies such as text infilling (Donahue et al., 2020), prompting (Liu et al., 2021), and QA (Jiang et al., 2021). PLMs imitate a more high-level information store, allowing for greater abstractness, flexibility, and generalizability. They are also able to better exploit contextual information than simple retrieval.

Studies have also shown that as PLMs scale up, they have have emergent abilities (Wei et al., 2022a), including reasoning. There has been increasing attention on their commonsense reasoning through works like COMET (Bosselut et al., 2019). However, studies show that even large PLMs struggle with commonsense tasks that humans can reason through very easily (Talmor et al., 2020). There are works that investigate more complicated reasoning tasks, e.g. arithmetic and symbolic reasoning (Wei et al., 2022b). PLMs inherently have some extent of reasoning capability, and many more complex reasoning tasks are easier to carry out over abstract PLM embedding space.

Template	Full Text with Explanation
A person with *Costochondritis* has a/an *exercise risk factor* because/since/as *{explanation}*	A person with Costochondritis has an exercise risk factor because *costochondritis can be aggravated by any activity that places stress on your chest area.*
A person with *gout* has a/an *lose weight prevention* because/since/as *{explanation}*	A person with gout has a lose weight prevention because *losing weight can lower uric acid levels in your body and significantly reduce the chance of gout attacks.*
A person with *rheumatoid* has a/an *therapy treatment* because/since/as *{explanation}*	A person with rheumatoid has a therapy treatment because *physiotherapy helps rheumatoid patients with pain control, reducing inflammation and joint stiffness and to return to the normal activities of daily living or sports.*

Table 1: Examples of **CHARD** templates with explanations (from CHARDAT). The human was asked to write the entire output text (not just the explanation) by infilling the template.

In this paper, we are interested in the intersection of these areas. Can PLMs act as knowledge bases and also reliably reason using their own knowledge? We investigate whether PLMs can learn and reason through health-related knowledge. Work on generation-based reasoning for health has been limited, with most prior work exploring retrieval-based methods. Generation-based reasoning is more difficult, as such a specialized domain contains esoteric information not prevalent in the PLM's training data, and involves a higher degree of specialized reasoning to handle domain-specific problems.

Healthcare is an important domain that deals with human lives. It is a large application area for machine learning and NLP. The need for automation in healthcare rises, as countless studies show that healthcare workers are overworked and burned out, especially recently due to the COVID-19 pandemic (Portoghese et al., 2014; Brophy et al.,

* Work done while at CMU.

Code: https://github.com/styfeng/CHARD

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 313–327
May 2-6, 2023

2021; Couarraze et al., 2021). Further, healthcare resources will continue to be strained as the baby boomer generation ages (Canizares et al., 2016).

To this end, we propose **CHARD**: Clinical Health-Aware Reasoning across Dimensions (§2.1). This task is designed to explore the capability of text generation models to act as implicit clinical knowledge bases and generate textual explanations about health-related conditions across several dimensions. The ultimate goal of **CHARD** is to eventually have a model that is knowledgeable and insightful across numerous clinical dimensions and reasoning pathways. For now, we focus on three relevant clinical dimensions using a template infilling approach, and collect an associated dataset, CHARDAT, which includes information for 52 health conditions across these dimensions (§2.2).

We perform extensive experiments on CHARDAT using two SOTA seq2seq models: BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) (§3.1), with data augmentation using backtranslation (Sennrich et al., 2016) (§3.2,4.2). We benchmark our models through automatic, human, and qualitative analyses (§5). We show that our models show strong potential, but have room to improve, and that **CHARD** is highly challenging with room for additional exploration. Lastly, we discuss several potential directions for improvement (§6).

2 Task and Dataset

2.1 The CHARD Task

Our task, **CHARD**: Clinical Health-Aware Reasoning across Dimensions, investigates the capability of text generation models to produce clinical explanations about various health conditions across several clinical dimensions (DIM). Essentially, we assess how a PLM can be used as and reason through an implicit clinical knowledge base.

We focus on three DIM: **risk factors** (RF), **treatment** (TREAT), and **prevention** (PREV), as they are important and relevant in the context of health. A risk factor refers to something that increases the chance of developing a condition. For cancer, some examples are age, family history, and smoking. Treatment refers to something that helps treat or cure a condition. For migraines, some examples are medication, stress management, and meditation. Prevention refers to strategies to stop or lower the chance of getting a condition. For diabetes, some examples are a healthy diet and regular exercise.

As an initial approach to **CHARD**, we use a template infilling formulation, where given an input template that lays out the structure of the desired explanation, the model's goal is to generate a complete explanation of how the particular DIM attribute relates to the given condition. In particular, the templates end with an {explanation} span that the models fill in by explaining the appropriate relationship. Some examples are in Table 1.

2.2 CHARDAT Dataset

Collection Process: We collect a dataset for our task called CHARDAT (where DAT is short for data). We collect data across the three DIM for 52 health conditions, listed in Appendix A. This is a manually curated list of health conditions which range from common conditions such as migraine and acne to rare conditions such as Lyme disease and Paget–Schroetter. The conditions were also selected by volume of online activity (e.g. number of active subreddit users), treatable vs. chronic conditions, and whether a condition can be self-diagnosed or not. This allows us to assess **CHARD** across a variety of conditions.

For each DIM, we manually collect an exhaustive list of DIM-related attributes (e.g. risk factors) for each condition. By *attribute*, we refer to a particular example of that DIM (e.g. "obesity"). This was accomplished by searching through reliable and reputable medically-reviewed sources such as MayoClinic, CDC, WebMD, and Healthline.

We collect the final text (with explanations) using Amazon Mechanical Turk (AMT). We ask approved AMT workers (with strong qualifications and approval ratings on healthcare-related tasks) to write factually accurate, informative, and relatively concise passages given a particular condition and DIM attribute template (per HIT), while encouraging them to consult the aforementioned health resources. Three separate annotation studies (one per DIM) with strict quality control were conducted to collect an annotation per example.[1] Annotations were regularly verified by authors, and a large subset of CHARDAT was manually examined for medical accuracy. More details are in Appendix B. Some examples from CHARDAT are in Table 1.

Splits and Statistics: We split CHARDAT by DIM into train, val, and test splits of ≈ 70%/15%/15%, and combine the individual splits per DIM to form the final splits called CHAR-

[1]Explanations for **CHARD** are typically quite standardized, and additional annotations were repetitive. Differences are mainly in language, so we instead opt for paraphrasing data augmentation techniques such as backtranslation (§3.2).

Dataset Stats	Train	Val	Test (seen/unseen)
# conditions = 52	44	39	41 (37/4)
RF = 52	44	26	26 (22/4)
TREAT = 52	43	21	20 (16/4)
PREV = 44	35	11	21 (17/4)
# sentences = 937	655	141	141 (70/71)
RF = 457	319	69	69 (32/37)
TREAT = 297	207	45	45 (20/25)
PREV = 183	129	27	27 (18/9)
Avg length = 36.2	37.7	36.1	35 (35.9/34.2)

Table 2: CHARDAT statistics. Differing #s by DIM are because there are more risk factors for most conditions, and some do not have prevention strategies. Length is in words.

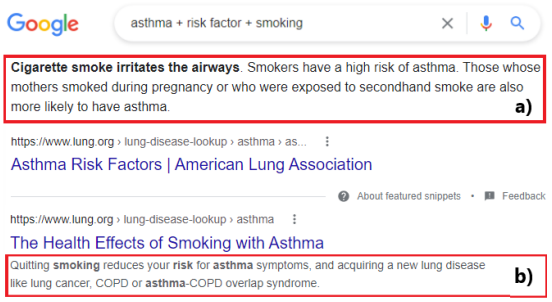

Figure 1: An example of the Google search results for the query {*asthma* + *risk factor* + *smoking*} highlighting: a) the *featured snippet*, b) the text below the first link.

DAT$_{tr}$, CHARDAT$_{val}$, and CHARDAT$_{test}$, respectively. The individual DIM splits are called DIM$_{tr}$, DIM$_{val}$, and DIM$_{test}$, where DIM is a short-form of the particular dimension: RF, TREAT, or PREV. The individual dimension subsets of CHARDAT are called CHARDAT$_{DIM}$.

For each DIM's test split, we ensure that approximately half consist of examples from conditions entirely unseen during training for that DIM, called DIM$_{test-unseen}$. This is to assess whether the model can generalize to unseen conditions. The other half contains examples from conditions seen during training called DIM$_{test-seen}$, but the specific condition and DIM attribute combination was unseen. The combined halves (across DIM) are called CHARDAT$_{test-unseen}$ and CHARDAT$_{test-seen}$. We do the same for the val split to ensure consistency for model selection purposes. CHARDAT statistics are in Table 2.

3 Methodology

3.1 Models

BART and T5: We experiment using two pretrained seq2seq models: BART and T5 (both base and large versions). These are suitable for our task formulation (template infilling). T5 (Raffel et al., 2020) has strong multitask pretraining. BART (Lewis et al., 2020) is trained to reconstruct original text from noised text (as a denoising autoencoder). We use their HuggingFace codebases.

Retrieval Baseline (RETR): We use a retrieval-based approach as a baseline. We manually query Google using {*condition* + DIM + DIM *attribute*}, e.g. {*asthma* + *risk factor* + *smoking*}, and extract either the *featured snippet* at the top of the results page, or the text below the first link if there is no featured snippet. If the featured snippet is a list or table, we manually concatenate the items into a single piece of text. An example is in Figure 1.

The extracted text approximates an explanation, which we then concatenate to the first part of the associated template to form the final text, e.g. *A person with asthma has a/an smoking risk factor because/since/as* ***{retrieved explanation}***. RETR leverages Google's strong search and summarization capabilities, serving as a useful baseline. Further, Google Search is an evolving baseline that continually challenges our **CHARD** models.[2]

3.2 Data Augmentation (DA)

Since CHARDAT is relatively small, which is mainly a function of our task and domain, i.e. there are a limited number of non-obscure medical conditions and associated DIM attributes, we hypothesize that data augmentation (DA) techniques (Feng et al., 2021a, 2020) may be useful.

As noted by Feng et al. (2021a), text generation and specialized domains (such as healthcare) both present several challenges for DA. In our case, many explanations contain clinical or health jargon which makes techniques that leverage lexical databases such as WordNet, e.g. synonym replacement (Feng et al., 2020), challenging or impossible.

We decide to use backtranslation (BT) (Sennrich et al., 2016) to augment examples in CHARDAT$_{tr}$, a popular and easy DA technique which translates a sentence into another language and back to the original language.[3] This usually results in a slightly altered version (paraphrase) of the original text. BT is effective here as healthcare-related terms are preserved relatively well, and the resulting paraphrased explanation remains relatively intact.

We use UDA (Xie et al., 2020) for BT, which translates sentences from English to French, then back to English. UDA is a DA method that uses unsupervised data through consistency training on $(x, DA(x))$ pairs. An advantage of UDA's BT is that we can control for the degree of variation using

[2]We will release our current baseline data.

[3]This is sometimes referred to as *round-trip translation*.

Tmp	Text
0	A person with acne has an avoid irritants prevention because *using oily or irritating personal care products clog your pores causing acne.*
0.5	*if you use oily or irritant personal care products, you block pores and cause acne.*
0.7	*using oily or irritating personal care products, you block acne pores.*
0.9	*use oily and irritating disinfectant products freezing your pores to cause the Acne restructurs.*
0	A person with MultipleSclerosis has a stress management prevention because *stress is more likely to exacerbate the symptoms of MS and bring about a flare or relapse.*
0.5	*stress is more likely to exacerbate MS symptoms and lead to an outbreak or relapse*
0.7	*stress is more likely to exacerbate symptoms of MS and trigger a flare or relapse.*
0.9	*severe mourning problems occurred at Vancouver Hospital (Prince Edward Island), British Columbia. (...)*

Table 3: Examples of original (tmp=0) and BT text. The explanation portion (which is backtranslated) is italicized.

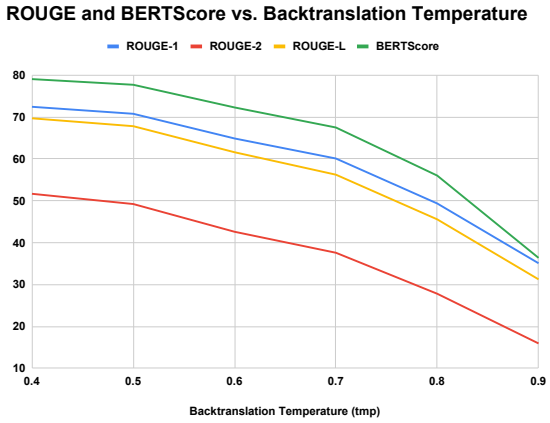

Figure 2: Graph showing how avg. ROUGE and BERTScore of BT vs. original text vary by BT tmp on CHARDAT_{tr}.

a *temperature (tmp)* parameter, where higher values (e.g. 0.9) result in more varied paraphrases. We only backtranslate the explanation portion of examples (concatenating them back to the preceding part) as we wish to keep the preceding part intact.

From the examples in Table 3, we can see that higher tmp typically results in more varied text, albeit with issues with content preservation and fluency. For the second example, the tmp=0.9 BT is completely unrelated to the original text. This is not entirely undesirable, as some noise may make our trained models more robust. From Figure 2, we see that the average ROUGE and BERTScore of backtranslated CHARDAT_{tr} text compared to the original text decrease as tmp increases, as expected.

3.3 Evaluation Metrics

We use several standard text generation evaluation metrics including reference-based token and semantic comparison metrics used in works like Lin et al. (2020) such as ROUGE (Lin and Hovy, 2003), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). SPICE translates text to semantic scene graphs and calculates an F-score over graph tuples. CIDEr captures sentence similarity, grammaticality, saliency, importance, and accuracy.[4]

We also use average word length (Len), BERTScore (Zhang et al., 2019), and Perplexity (PPL). BERTScore serves as a more semantic similarity measure by assessing BERT (Devlin et al., 2019) embeddings similarity between individual tokens. We multiply by 100 when reporting BERTScore. PPL approximately measures fluency, where lower values represent higher fluency. We use GPT-2 (Radford et al., 2019) for PPL. Higher is better for all metrics other than PPL and Len.

4 Experimental Setup

4.1 Model Finetuning and Generation

For the standard (non-augmented) **CHARD** models, we train and evaluate four versions of each on CHARDAT, CHARDAT_{RF}, CHARDAT_{TREAT}, and CHARDAT_{PREV}, respectively. The first of these is a combined model that learns to handle all three DIM at once depending on the DIM given at inference, while the latter three are models trained on each individual DIM. We predict that while the latter three may perform better on their particular DIM, the first model is more effective overall as it accomplishes our goal of having a single PLM that can store knowledge and reason through several DIM. It is thus more adaptable and generalizable.

For training the **CHARD** models, we keep most hyperparameters static, other than learning rate (LR) which is tuned per individual model. For each model, we select the epoch that corresponds to highest ROUGE-2 on CHARDAT_{val}, and decode using beam search. See Appendix C for more.

4.2 Data Augmentation Experiments

We try several backtranslation DA experiments.

2x DA with Different Tmp: Our first set of experiments involves 2x DA (backtranslating each CHARDAT_{tr} explanation once, to 2x the original training data) using different BT tmp which we call BT-set: {0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. We predict that the optimal tmp lies in the 0.6-0.7 range, as the text is modified to a reasonable degree.

Different DA Amounts (2x-10x): We also try further DA amounts: 3x, 4x, 5x, 7x, and 10x the original amount of training data. We explore whether the amount of augmentation affects performance, and hypothesize that performance will

[4]Matching metrics are sufficient as **CHARD** explanations are standardized (space for explanations is low) since our inputs present a particular condition and DIM attribute combo.

increase to a certain point and decline afterward. This is because the advantages of DA may taper off since the augmented data are variations of the original, and models may overfit past a point.

DA Amount Strategies (best-tmp vs. diff-tmp): We also investigate two strategies for selecting each successive iteration of augmented examples. The first is best-tmp, where all the augmented data comes from BT of the tmp that performed best for 2x DA (e.g. all from 0.7).[5]

The second is diff-tmp, where each successive iteration is the tmp that performed next best (e.g. 2x is the best tmp, 3x is additionally the second-best tmp, etc.). For the highest DA amounts (e.g. 10x), when the six tmp values in BT-set have been exhausted, we go back to the best tmp and repeat.

Base vs. Large Models: For the base models (BART-base and T5-base), we try all aforementioned tmp, DA amounts, and amount strategies. For the large models, we try the top three temperatures (for 2x DA) and amount strategy that performed best on the corresponding base model, and only 3x, 5x, 7x, and 10x DA amounts.

Note that BT tmp and DA amounts are both hyperparameters, so while we train models corresponding to different values of them, the final chosen models correspond to the ones that performed best on CHARDAT$_{val}$. We then use these final models to generate on CHARDAT$_{test}$. We report the results of the overall best models in §5.

4.3 Human Evaluation

We conduct human evaluation using AMT.[6] We ask two approved annotators (with strong qualifications and approval ratings on healthcare-related tasks) to each evaluate all 141 CHARDAT$_{test}$ examples. Our evaluation uses pairwise comparison of the outputs from two methods, split into three studies per DIM: RETR vs. best **CHARD** model, RETR vs. human, and human vs. best **CHARD** model.

We ask annotators to choose which amongst the two outputs (presented in a random order per example) has better 1) medical accuracy (MedAcc), 2) informativeness (Inform), and 3) readability (Read). Medical accuracy refers to which explanation is more clinically correct for the given DIM attribute and condition. Informativeness refers to which is more complete and explains in sufficient detail (including *why?*). Readability refers to which is more readable, which includes fluency (natural-sounding English) and conciseness/brevity (not overly long).

There are 3 choices for each evaluation aspect - O1: first is better, O2: second is better, O3: both are indistinguishable. To aggregate multiple annotations per example, we find the overall fraction of responses towards each outcome value.

[5]This is possible because UDA uses sampling, so even for the same tmp, the backtranslations differ each time.

[6]See Appendix D for further human evaluation details.

Metric	RETR	BART-base	BART-large	T5-base	T5-large
ROUGE-1	43.30	51.37	**51.54**	50.00	50.66
ROUGE-2	28.18	39.35	**40.27**	38.31	37.74
ROUGE-L	39.03	49.55	**49.88**	48.07	48.05
BLEU-1	32.20	31.20	28.40	32.60	**34.30**
BLEU-2	25.20	27.10	24.90	28.10	**29.20**
BLEU-3	21.50	24.70	22.90	25.50	**26.40**
BLEU-4	18.50	23.00	21.30	23.60	**24.30**
METEOR	**24.40**	22.10	22.10	21.80	22.10
CIDEr	2.36	8.56	6.90	8.71	**9.03**
SPICE	35.10	50.50	**50.70**	49.10	49.20
BERTScore	39.54	60.04	**60.78**	59.80	59.00
PPL	65.27	61.00	87.45	56.78	**52.52**
Len	52.80	20.16	18.60	21.35	22.23

Table 4: Avg. auto eval results of RETR and the best models (for BART and T5) on CHARDAT$_{test}$. Bold corresponds to best performance. For human text, PPL = 67.86, Len = 35.04.

Metric	test split (full)	test-seen	test-unseen
ROUGE-1	50.66	49.42	**51.93**
ROUGE-2	37.74	37.04	**38.35**
ROUGE-L	48.05	46.98	**49.12**
BLEU-1	34.30	33.50	**35.20**
BLEU-2	29.20	28.60	**29.90**
BLEU-3	26.40	25.90	**27.00**
BLEU-4	24.30	23.80	**24.80**
METEOR	22.10	21.60	**22.60**
CIDEr	9.03	**10.31**	7.59
SPICE	49.20	48.60	**49.80**
BERTScore	59.00	57.79	**60.18**
PPL	52.52	**51.06**	53.96
Len	22.23	22.73	21.73

Table 5: Avg. auto eval results of T5-large on CHARDAT$_{test}$ and the test-seen and test-unseen halves.

5 Results and Analysis

We report automatic results on CHARDAT$_{test}$ of the best models (for BART-base, BART-large, T5-base, T5-large) trained on CHARDAT compared to RETR in Table 4. The best models are tmp=0.9 2x DA for BART (base and large), 5x DA with diff-tmp for T5-base, and tmp=0.6 2x DA for T5-large.

Our best overall **CHARD** model is T5-large based on automatic results and qualitative analysis. We break down results of T5-large on CHARDAT$_{test-seen}$ and CHARDAT$_{test-unseen}$ in Table 5. We show results of T5-large compared to T5-large$_{DIM}$ (models trained on the individual DIM) in Table 6. We conduct human evaluation with T5-large, and the results are in Table 7.

Graphs displaying models' ROUGE-2 on CHARDAT$_{val}$ for 2x DA across various BT tmp and different DA amounts can be found in Figures 3 and 4, respectively. Tables 8 and 9 contain qualitative examples, with more in Appendix E.

	Risk Factors (RF_{test})		Treatment ($TREAT_{test}$)		Prevention ($PREV_{test}$)	
Metric	**T5-large**	**T5-large$_{RF}$**	**T5-large**	**T5-large$_{TREAT}$**	**T5-large**	**T5-large$_{PREV}$**
ROUGE-1	52.74	**53.17**	**49.42**	47.38	47.73	**50.00**
ROUGE-2	40.52	**41.88**	36.12	**36.69**	33.00	**36.10**
ROUGE-L	50.40	**51.03**	**46.60**	45.54	44.43	**48.19**
BLEU-1	**34.80**	34.70	**31.10**	25.70	**30.80**	29.80
BLEU-2	30.40	**30.90**	**26.30**	22.40	**25.10**	25.00
BLEU-3	27.90	**28.60**	**23.90**	20.70	21.90	**22.00**
BLEU-4	26.10	**26.90**	**22.10**	19.30	**19.30**	**19.30**
METEOR	23.00	**24.20**	**20.70**	19.20	20.50	**20.80**
CIDEr	**13.50**	10.57	5.06	**5.98**	**5.88**	5.83
SPICE	49.90	**51.50**	**46.60**	45.30	46.50	**46.60**
BERTScore	60.40	**61.03**	**58.07**	56.60	56.90	**59.09**
PPL	**40.90**	58.92	**52.13**	86.15	**84.06**	110.66
Len	22.30	20.28	22.09	19.82	22.27	20.52

Table 6: Breakdown of the avg. auto eval results of T5-large compared to T5-large$_{DIM}$ models (trained on the three individual DIM) on the respective DIM subsets of CHARDAT$_{test}$. Bold corresponds to best performance per DIM.

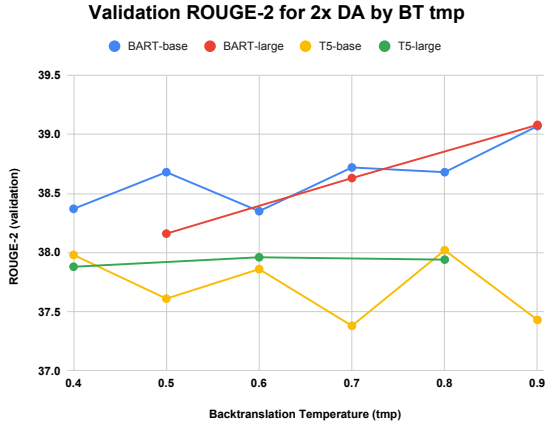

Figure 3: Graph showing how avg. ROUGE-2 on CHARDAT$_{val}$ varies by backtranslation temperature for 2x DA.

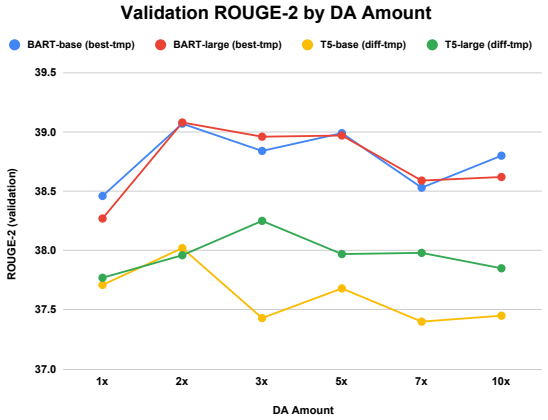

Figure 4: Graph showing how avg. ROUGE-2 on CHARDAT$_{val}$ varies by DA amount. 1x essentially refers to no DA.

Methods	Aspect	O1	O2	O3
RETR vs. Human	MedAcc	0.45	**0.53**	0.02
	Inform	0.45	**0.53**	0.02
	Read	0.22	**0.69**	0.09
Human vs. T5	MedAcc	**0.72**	0.24	0.04
	Inform	**0.72**	0.25	0.03
	Read	0.41	**0.49**	0.10
RETR vs. T5	MedAcc	**0.73**	0.25	0.02
	Inform	**0.73**	0.26	0.01
	Read	0.35	**0.62**	0.03

Table 7: Avg. human eval results on CHARDAT$_{test}$. O1: first method wins, O2: second method wins, O3: indistinguishable. Bold corresponds to higher fractional outcome. T5 refers to T5-large. See §4.3 and Appendix D for further details.

Method	**Text**
Input (1)	A person with Hyperhidrosis has a medication treatment because {explanation}
Human	A person with Hyperhidrosis has a medication treatment because anticholinergics medicines such as glycopyrrolate, oxybutynin, benztropine, propantheline etc are helping to prevent the stimulation of sweat glands and, thus, decreasing sweating.
RETR	A person with Hyperhidrosis has a medication treatment because the most common medications used to manage excessive sweating are anticholinergics, including but not limited to: Glycopyrrolate (Robinul). Oxybutynin (Ditropan). Benztropine. Propantheline bromide.
T5-L	A person with Hyperhidrosis has a medication treatment because medications such as ibuprofen, phenytoin, and phenobarbital can help reduce the amount of sweating in the body.
T5-B	A person with Hyperhidrosis has a medication treatment because it helps to reduce the amount of blood in the body and reduces the risk of a blood clot.
Input (2)	A person with Alzheimers has an engage in activity prevention because {explanation}
Human	A person with Alzheimers has an engage in activity prevention because engaging in activity will increase nerve connections in the brain or release hormones that increase nerve cell creation and this helps preserve nerve cells in the brain and mental function
T5-L	A person with Alzheimers has an engage in activity prevention because activities such as walking, jogging, swimming, and playing games can help prevent Alzheimer's
BART-B	A person with Alzheimers has an engage in activity prevention because a person with Alzheimer's disease is more likely to develop dementia
Input (3)	A person with Prostatitis has an infection risk factor because {explanation}
Human	A person with Prostatitis has an infection risk factor because the condition is caused by common strains of bacteria which spread from other parts of the urinary or reproductive systems to infect and inflame the prostate resulting in prostatitis.
T5-L	A person with Prostatitis has an infection risk factor because infections of the prostate, urethra, and genital tract can lead to prostatitis.

Table 8: Qual examples for CHARDAT$_{test-seen}$: human, RETR, T5-large (L), T5-base (B), BART-base (B).

5.1 Automatic Evaluation Results

From Table 4, we see that all **CHARD** models perform better than RETR across most metrics. RETR's average outputs are much longer than those of our models and humans. Among our models, T5-large and BART-large perform best, demonstrating that larger models are more adept. T5-large performs best overall (combined with the qual analysis in §5.3), with the longest average outputs among our models. Some of our models achieve better average fluency (PPL) compared to humans, but the outputs are generally noticeably shorter.

From Table 5, we see that T5-large surprisingly performs better on the test-unseen half. It appears that the model can generalize decently to unseen conditions when trained on CHARDAT. This may partially be due to similar explanations for particular DIM attributes across conditions, e.g. why *sleep* helps treat some conditions may be similar.

From Table 6, we see that for most DIM (namely RF and PREV), the model trained on that specific DIM performs better on that DIM. However, our

Method	Text
Input (4)	A person with Costochondritis has a fix improper posture prevention because {explanation}
Human	A person with Costochondritis has an fix improper posture prevention because bad posture puts stress on joints and tissues, causing joint stiffness, weakened muscles, and inflammation that is so common in Costochondritis.
RETR	A person with Costochondritis has a fix improper posture prevention because what can I do to relieve costochondritis? Lifestyle and home remedies. Nonprescription pain relievers. Acetaminophen (Tylenol, others) ibuprofen (Advil, Motrin IB, others) or naproxen sodium (Aleve, others) (...)
T5-L	A person with Costochondritis has a fix improper posture prevention because poor posture can lead to pain and stiffness.
Input (5)	A person with rheumatoid has a diet risk factor because {explanation}
Human	A person with rheumatoid has a diet risk factor because unhealthy food choices can increase inflammation from rheumatoid
RETR	A person with rheumatoid has a diet risk factor because the results of this study revealed that high intakes of carbohydrates (e.g., potatoes) might be associated with increased RA risks.
T5-L	A person with rheumatoid has a diet risk factor because people who eat a high-fat diet are at a higher risk of developing the disease.
BART-L	A person with rheumatoid has a diet risk factor because a diet rich in red meat, dairy products, and dairy products can cause inflammation
Input (6)	A person with depression has a life-changing event risk factor because {explanation}
Human	A person with depression has a life-changing event risk factor because approximately 70% of first depression episodes and 40% of recurrent episodes of depression are preceded by a severe life event
RETR	A person with depression has a life-changing event risk factor because psychological and Psychosocial Risk Factors for Recurrence · by SL Burcusa · 2007 · Cited by 1085 — These factors include age at onset of the first episode, lifetime number of
T5-L	A person with depression has a life-changing event risk factor because a major life event, such as the death of a loved one, can increase the risk of depression.

Table 9: Qual examples for CHARDAT$_{test-unseen}$: human, RETR, T5-large (L), BART-large (L).

general T5-large model performs better on TREAT. It may be that training on CHARDAT has allowed the model to learn from data of other DIM, improving its overall knowledge and generation capabilities (an advantage of a single combined model).

From Figure 3, we see that the BART models generally increase in performance with higher BT tmp (upward trend), whereas T5 fluctuates. This may be due several reasons, e.g. differences in the architecture and pretraining strategies of the models, allowing BART to leverage noisy data more effectively. From Figure 4, we see that performance generally increases for each model up to a certain point (e.g. 2x or 3x DA), and then decreases afterward, aligning with our hypothesis from §4.2.

5.2 Human Evaluation Results

From Table 7, we see that both RETR and T5-large are outperformed by humans, although RETR is relatively close in informativeness and medical accuracy, and T5-large slightly outperforms on readability. RETR outperforms T5-large on medical accuracy and informativeness, which is somewhat expected as it uses Google Search. It is worse than T5-large on readability, as our models generate more fluent, concise, and readable text (see §5.3).

5.3 Qualitative Analysis

We examine the qualitative examples in Tables 8 and 9. Firstly, we see that RETR is able to generally perform well by extracting relevant information (ex.1 - a list of medications for Hyperhidrosis, ex.5 - that carbohydrates increase RA risk), which is expected using Google Search. However, it sometimes extracts a lengthy amount of irrelevant information. For ex.4, RETR extracts a difficult-to-read list of different TREAT strategies, which is for the wrong DIM, and does not narrow down on an explanation for the specific DIM attribute in the input. For ex.6, it extracts the info and beginning of a passage from a scientific article, ending abruptly and not explaining the given DIM attribute.

Our models, specifically T5-large, are generally able to output more concise, readable, and sometimes relevant explanations compared to RETR. For ex.1, T5-large outputs a list of medications, albeit not for Hyperhidrosis - showing weaknesses in medical accuracy. Other than ibuprofen, the other medications are not in CHARDAT$_{tr}$, showing that these were likely already known to T5-large through pretraining. For ex.2, it generates a reasonable list of activities to help prevent Alzheimer's, and for ex.3, it lists correct body parts where an infection can occur to cause Prostatitis. It can generalize well to unseen conditions, as shown through ex.4-6. It reasons that poor posture can lead to pain and stiffness, high-fat diets can increase the chance of rheumatoid, and that a major life event (*"death of a loved one"*) can cause depression. These generalization capabilities are likely from a combination of pretraining and CHARDAT$_{tr}$.

Compared to humans, T5-large's outputs are lacking. Human explanations are typically longer and more informative, explaining the exact reason (*why?*) a specific DIM attribute relates to the given condition. For ex.2, it explains how activities can help *"preserve nerve cells in the brain and mental function"*, whereas T5-large simply lists activities. This similarly occurs for ex.3-5. Human explanations are also typically more medically accurate, e.g. for ex.1, the listed medications are correct. However, we do see that some of T5-large's outputs (for ex.1,2,4) are more readable. Further, T5-large sometimes presents more information, e.g. an exact list of activities for ex.2, a specific type of diet (*"high-fat"*) for ex.5 (human just says *"unhealthy"*), and an example of a life-changing event for ex.6.

BART-large also performs decently. In ex.5, it lists several specific and correct types of foods (*"red meat, dairy products"*). The base models perform much worse. For ex.1, T5-base talks about medication reducing *"blood clots"*, unrelated to Hyperhydrosis. For ex.2, BART-base writes an explanation completely irrelevant to the input DIM.

6 Directions for Improvement

We see that our models are decent and generate readable text, but can improve on medical accuracy and informativeness. While they are not nearly ready for real-world use, they show potential.

As stated, the purpose of **CHARD** is to assess the capabilities of PLMs to act as implicit clinical knowledge bases that can reason through several dimensions. How can we improve our models, and possibly our dataset and task formulation?

Dataset and task formulation: We introduce **CHARD** and initially frame the task using a template infilling approach which is more constrained. More flexible formulations may better leverage the knowledge and generation capabilities of PLMs.

Our current approach involves generating explanations about a single condition and DIM attribute at a time. We can possibly improve CHARDAT by annotating for more complicated input queries. This is because a PLM may be more effective at answering more complicated queries, e.g. comparing and contrasting conditions and DIM and multi-hop reasoning. It is likely easier to make complicated inferences and connections over the abstract PLM embedding space than over retrieved text passages.

Further, we can expand CHARDAT to include more dimensions and topics in the health domain. These improvements may allow for the training of a single system that is able to make complicated clinical inferences across various topics and DIM.

Model improvements: We can explore models such as GPT-3 (Brown et al., 2020) and PALM (Chowdhery et al., 2022) for **CHARD** that are larger with stronger pretraining. We can also investigate enhancing PLMs with information retrieval, e.g. using a retrieval approach to obtain relevant scientific literature as evidence, combined with a text summarization system to digest the content. Our model can then conduct its clinical reasoning on this digested content. Users can potentially take advantage of such a system to automatically verify the medical accuracy of generated explanations, and then improve the generation model itself using this feedback loop (i.e. a self-improving system).

7 Related Work

Constrained Text Generation: There have been several works on constrained text generation. For creative text generation, Gangal et al. (2022) introduce narrative reordering (NAREOR) to edit the temporality of narratives. Keh et al. (2022) and Keh et al. (2023) explore the generation of personifications and tongue twisters, respectively. Donahue et al. (2020) introduce and investigate the task of infilling. Feng et al. (2019) propose Semantic Text Exchange to adjust topic-level text semantics using infilling. Rajagopal et al. (2021) investigate cross-domain reasoning using a prompt-tuning setup. Our work distinctly investigates template infilling for clinical reasoning along dimensions.

Commonsense Reasoning for Models: One large commonsense KG is COMET, which trains on KG edges to learn connections between words and phrases. COSMIC (Ghosal et al., 2020) uses COMET to inject commonsense into models. CommonGen (Lin et al., 2020) assesses the commonsense reasoning of text generation models. Several works investigate CommonGen, including SAPPHIRE (Feng et al., 2021b) and VisCTG (Feng et al., 2022), the latter of which uses visual grounding. Unlike these works, **CHARD** distinctly investigates reasoning for the clinical/health domain.

Reasoning for Clinical/Health Domain: Most existing work here involves retrieval or extraction. MIMICause (Khetan et al., 2022) extracts causal medical information from electronic health records to help understand narratives in clinical texts. Ahne et al. (2022) extract a causal graph and reason about diabetes distress for better understanding the opinions, feelings, and observations of the diabetes online community from a causality perspective.

For generation, Moramarco et al. (2021) investigate the use of LMs to simplify medical text. Abaho et al. (2022) probe factual knowledge from LMs to elicit answers related to treatment outcomes. **CHARD** has a different goal: rather than simply probe for factual knowledge, we assess how LMs can act as and reason through an implicit knowledge base. Meng et al. (2022) investigate probing biomedical knowledge by introducing a benchmark, MedLAMA, that focuses on 19 relations. **CHARD** instead focuses on clinical knowledge reasoning along different dimensions.

8 Conclusion and Future Work

In conclusion, we proposed and investigated the task of **CHARD**: Clinical Health-Aware Reasoning across Dimensions, to explore the capability of text generation models to act as implicit clinical knowledge bases and generate explanations across several health dimensions. We presented a dataset, CHARDAT, and conducted experiments

with BART and T5. Extensive evaluation and qualitative analysis demonstrated that our models are decent, especially for generating concise and readable text, but can be improved on medical accuracy and informativeness, and that **CHARD** is challenging with much potential for further exploration. We highly encourage the research community to further investigate and improve upon **CHARD**.

Future directions are discussed in §6. Some additional ideas include trying more data augmentation strategies and decoding strategies for text infilling.

Limitations

We discuss some limitations of our work and potential directions for improvement in §6. Specifically, our template-infilling approach is less flexible, and we can expand to more complicated input queries to better leverage the power of PLMs in future work. Further, CHARDAT focuses on three main clinical dimensions, which can be expanded upon to include more dimensions and topics in the future. Our seq2seq models are also relatively weaker compared to GPT-3, PALM, and recent larger PLMs, which may perform more effectively on **CHARD**. We are also investigating a completely generative approach, and combining generation with retrieval in interesting ways may be more effective. Overall, our current **CHARD** models have room to improve on medical accuracy and informativeness, and are not nearly ready for real-world use.

However, we note again that we are the first to propose **CHARD**, and our work is the first step towards longer-term goals regarding clinical reasoning using PLMs. We are after more of the *commonsense* medical reasoning for now, rather than very deep medical knowledge. In this paper, we see how far one can get with a standard task formulation, NLP methods, seq2seq models, and AMT annotations. As they say, *"walk before you run"*!

Ethics

We collected CHARDAT and conducted our human evaluation studies using AMT, in a manner consistent with terms of use of any sources and intellectual property and privacy rights of AMT crowd workers.

Our collected dataset, CHARDAT, consists of general clinical information, where explanations are impersonal. We also manually examined a large subset of the data, and ensured there were no issues with respect to privacy and other ethical concerns, e.g. offensive words, profanities, racism, gender bias, and other malicious language.

We acknowledge the weaknesses of **CHARD** models and the potential risks if they are used for real-world purposes. We will never use our models or encourage their use for real-world purposes, at least in their current state, and also emphasize this in the paper. As we noted, we propose **CHARD** and conduct our initial experiments purely for investigation purposes and to test our hypotheses. Our paper presents an important contribution to the ML, NLP, and healthcare communities, and we encourage researchers to further improve upon it.

Our task, models, dataset, and accompanying publication are intended only for research purposes and to assess the capabilities of text generators.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2022. Position-based prompting for health outcome generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 26–36, Dublin, Ireland. Association for Computational Linguistics.

Adrian Ahne, Vivek Khetan, Xavier Tannier, Md Imbesat Hassan Rizvi, Thomas Czernichow, Francisco Orchard, Charline Bour, Andrew Fano, and Guy Fagherazzi. 2022. Extraction of explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets from 2017 to 2021: Deep learning approach. *JMIR Med Inform*, 10(7):e37201.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

James T. Brophy, Margaret M. Keith, Michael Hurley, and Jane E. McArthur. 2021. Sacrificed: Ontario healthcare workers in the time of covid-19. *NEW SOLUTIONS: A Journal of Environmental and Occupational Health Policy*, 30(4):267–281. PMID: 33174768.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Mayilee Canizares, Monique Gignac, Sheilah Hogg-Johnson, Richard Glazier, and Badley Elizabeth. 2016. Do baby boomers use more healthcare services than other generations? longitudinal trajectories of physician service use across five birth cohorts. *BMJ Open*, 6.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Sébastien Couarraze, Louis Delamarre, Fouad Marhar, Binh Quach, Jiao Jiao, Raimundo Avilés Dorlhiac, Foued Saadaoui, Andy Su-I Liu, Benoït Dubuis, Samuel Antunes, Nicolas Andant, Bruno Pereira, Ukadike C. Ugbolue, Julien S. Baker, The COVISTRESS network, Maëlys Clinchamps, and Frédéric Dutheil. 2021. The major worldwide stress of healthcare professionals during the first wave of the covid-19 pandemic – the international covistress survey. *PLOS ONE*, 16(10):1–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. GenAug: Data augmentation for finetuning text generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 29–42, Online. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021a. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Steven Y. Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy, and Varun Gangal. 2021b. SAPPHIRE: Approaches for enhanced concept-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 212–225, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Steven Y. Feng, Aaron W. Li, and Jesse Hoey. 2019. Keep calm and switch on! preserving sentiment and fluency in semantic text exchange. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2701–2711, Hong Kong, China. Association for Computational Linguistics.

Steven Y. Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. 2022. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10618–10626.

Varun Gangal, Steven Y. Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10645–10653.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Sedrick Scott Keh, Steven Y. Feng, Varun Gangal, Malihe Alikhani, and Eduard Hovy. 2023. Pancetta: Phoneme aware neural completion to elicit tongue twisters automatically. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*.

Sedrick Scott Keh, Kevin Lu, Varun Gangal, Steven Y. Feng, Harsh Jhamtani, Malihe Alikhani, and Eduard Hovy. 2022. PINEAPPLE: Personifying INanimate entities by acquiring parallel personification data for learning enhanced generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6270–6284, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Vivek Khetan, Md Imbesat Rizvi, Jessica Huber, Paige Bartusiak, Bogdan Sacaleanu, and Andrew Fano. 2022. MIMICause: Representation and automatic extraction of causal relation types from clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 764–773, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, Dublin, Ireland. Association for Computational Linguistics.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, and Nils Hammerla. 2021. Towards more patient friendly clinical notes through language models and ontologies.

Igor Portoghese, Maura Galletta, Ross Coppola, Gabriele Finco, and Marcello Campagna. 2014. Burnout and workload among health care workers: The moderating role of job control. *Safety and Health at Work*, 5:152–157.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Dheeraj Rajagopal, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, Andrew Fano, and Eduard Hovy. 2021. Cross-domain reasoning via template filling.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Full List of Health Conditions

See Table 10 for a list of all health conditions in CHARDAT.

B CHARDAT Annotation Details

Human annotation for CHARDAT was done via paid crowdworkers on AMT, who were from Anglophone countries. They were selected through a series of qualification tests on a small subset of the samples, and have a history of high approval rates ($> 95\%$) and good performance on related tasks. Based on initial annotations and performance on the qualification tests, workers were only re-hired if their performance was sufficient over time and they reliably followed the given instructions. The annotators were paid variable amounts (with periodic bonuses over time) depending on their performance and consistency, and the pay for all workers exceeds the minimum wage for the USA.

The workers were asked to write passages (that include explanations) that are as specific and factually accurate as possible, describing how a specific dimension attribute relates to the given condition. Each HIT (annotation page) contains a single condition + dimension attribute combination, and they write a single passage that fills in the given template with an explanation. In the instructions, we describe each dimension in detail, and include several examples of correct and incorrect passages (regarding medical/factual accuracy, brevity/readability, and informativeness). We also encourage them to consult useful and trusted clinical resources such as MayoClinic, CDC, WebMD, and Healthline, if necessary, while writing the explanation.

Annotations were manually examined by the authors as they came in, and annotators were asked to improve their explanations if necessary. Annotators with consistently poor annotations were asked to stop annotating, and their completed annotations were re-annotated by others. At the end of the data collection process, the authors manually examined a large subset of CHARDAT, ensuring sufficiently high quality of annotations in terms of medical accuracy, informativeness, and readability.

C Further Model Finetuning and Generation Details

T5-large consists of 770M params, T5-base 220M params, BART-large 406M params, and BART-base 139M params. For all models, we use beam search with a beam size of 5, decoder early stopping, a decoder length penalty of 0.6, and a decoder minimum length of 1. We set the maximum encoder and decoder lengths depending on values that can fit all examples in CHARDAT_{tr}, which ended up being 32 and 128 (for encoder and decoder, respectively) for the BART models, and 35 and 128 for the T5 models. Models are trained using fp16, and Adam optimizer with epsilon=1e-08. We use a training seed of 42 for all models, and a random seed of 42 for all other scripts that involved randomization. Decoding is done using beam search with a beam width of 5.

For model training, we use a batch size of either 64 or 32 for T5-base and BART-base, and either 8 or 16 for BART and T5-large (depending on GPU memory). For T5-base and BART-base, we use 400 warmup steps, 500 for BART-large, and 1200 for T5-large. We train all models up to a reasonable number of epochs (e.g. 20 to 30 for base models and 10 to 15 for large models). The learning rates for **CHARD** models were determined by trying a range of values (e.g. from 1e-8 to 5e-1), and finding ones which led to good convergence behavior. For the best-performing models, learning rates are as follows: BART-base = 5e-06, BART-large = 1e-05, T5-base = 1e-03, T5-large = 1e-05.

Training was done using single GTX 1080 Ti, TITAN RTX, RTX 2080 Ti, and GTX TITAN X GPUs. Model training time varied depending on the model type+size and amount of data augmentation, and varied between 5 minutes to 3 hours.

D Further Human Evaluation Details

Human evaluation was done via paid crowdworkers on AMT, who were from Anglophone countries. They were selected through qualification tests and have a history of high approval rates ($> 95\%$) and good performance on related tasks. Each example was evaluated by 2 annotators. The time given for each AMT task instance or HIT was 1 hour maximum for an approximately 1-minute task. Sufficient time to read instructions, as calibrated by authors, was also considered. Annotators were paid 20 cents per HIT. This rate ($12/hr) exceeds the minimum wage for the USA ($7.25/hr) and constitutes fair pay. Workers who performed well were also paid periodic bonuses based on the timeliness and quality of their annotations.

The human evaluation was split into 9 studies: 3 pairwise method comparisons (RETR vs. T5-large,

Dysthymia	cfs	ibs	Narcolepsy	bulimia
Hypothyroidism	Costochondritis	psychosis	CysticFibrosis	POTS
MultipleSclerosis	Gastroparesis	gout	adhd	diabetes
CrohnsDisease	lupus	rheumatoid	Sinusitis	thyroidcancer
Hyperhidrosis	gerd	AnkylosingSpondylitis	endometriosis	schizophrenia
asthma	bipolar	depression	pcos	covid19
acne	anxiety	dementia	ptsd	dystonia
Epilepsy	ErectileDysfunction	Herpes	insomnia	Anemia
LymeDisease	migraine	ocd	parkinsons	Alzheimers
hpv	Prostatitis	backpain	Sciatica	Fibromyalgia
bpd	PagetSchroetter			

Table 10: A list of all 52 health conditions used in CHARDAT.

RETR vs. human, and human vs. T5-large) by 3 dimensions (risk factors, treatment, and prevention). Each HIT or task page presented a given condition, the associated dimension attribute, and two explanations (from the two methods of the study) in a random order. They are asked to select among the three choices (first explanation is better, second explanation is better, *hard to prefer one over the other*) for the three evaluation aspects of medical accuracy, informativeness, and readability. In the instructions, we describe the clinical dimension and each evaluation aspect in detail with positive and negative examples of each, and encouraged them to consult useful and trusted clinical resources such as MayoClinic, CDC, WebMD, and Healthline, if necessary, while evaluating the explanations.

E Further Qualitative Examples

See Tables 11 and 12 for more qualitative examples on CHARDAT$_{test}$.

Method	Text
Input	A person with Hyperhidrosis has a/an infection risk factor because/since/as {explanation}
Human	A person with Hyperhidrosis has an infection risk factor because some kinds of infections like tuberculosis, HIV, bone infection cause hyperhidrosis
RETR	A person with Hyperhidrosis has a/an infection risk factor because/since/as this increases the risk of skin conditions such as athlete's foot and more severe conditions such as bacterial infections or pitted keratolysis. Study results report a nearly 30% greater risk of skin infections in patients with hyperhidrosis compared with healthy controls.
T5-L	A person with Hyperhidrosis has an infection risk factor because infections such as staph, typhoid, hepatitis and polio can cause excessive sweating.
T5-B	A person with Hyperhidrosis has an infection risk factor because a person who has a viral infection is more likely to have a virus infection.
BART-L	A person with Hyperhidrosis has an infection risk factor because infection can cause hyperhidrosis
BART-B	A person with Hyperhidrosis has an infection risk factor because a person with hyperhidrosis is more likely to have an infection.
Input	A person with AnkylosingSpondylitis has a/an age risk factor because/since/as {explanation}
Human	A person with AnkylosingSpondylitis has an age risk factor because about 80 percent of people first develop ankylosing spondylitis symptoms under age 30
RETR	A person with AnkylosingSpondylitis has a/an age risk factor because/since/as about 80 percent of people first develop ankylosing spondylitis symptoms under age 30
T5-L	A person with AnkylosingSpondylitis has an age risk factor because people over 50 years of age are at a higher risk of developing the disease
T5-B	A person with AnkylosingSpondylitis has an age risk factor because older adults are more likely to develop ankylosis.
BART-L	A person with AnkylosingSpondylitis has an age risk factor because the risk of developing ankylosis increases with age
BART-B	A person with AnkylosingSpondylitis has an age risk factor because older people are more likely to develop the disease
Input	A person with rheumatoid has a/an regular exercise treatment because/since/as {explanation}
Human	A person with rheumatoid has a/an regular exercise treatment because it helps reduce the disability often associated with rheumatoid arthritis and boosts strength and flexibility
RETR	A person with rheumatoid has a/an regular exercise treatment because/since/as but exercise is one of the key treatments to help reduce the disability often associated with rheumatoid arthritis. Regular exercise can boost strength and flexibility in people who have rheumatoid arthritis. Stronger muscles can better support your joints, while improved flexibility can aid joint function.
T5-L	A person with rheumatoid has a regular exercise treatment because exercise can help reduce inflammation and improve blood flow to the joints.
T5-B	A person with rheumatoid has a regular exercise treatment because regular exercise can help reduce the symptoms of the disease.
BART-L	A person with rheumatoid has a regular exercise treatment because exercise can help reduce inflammation and pain
BART-B	A person with rheumatoid has a regular exercise treatment because regular exercise can help reduce inflammation and inflammation in the body.

Table 11: More qual examples for CHARDAT$_{test}$: human, RETR, T5-large (L), T5-base (B), BART-large (L), BART-base (B)

Method	Text
Input	A person with thyroidcancer has a/an chemotherapy treatment because/since/as {explanation}
Human	A person with thyroidcancer has a chemotherapy treatment because Chemotherapy (chemo) uses anti-cancer drugs that are injected into a vein or are taken by mouth. Chemotherapy is systemic therapy, which means that the drug enters the bloodstream and travels throughout the body to reach and destroy cancer cells and stops cancer growth.
RETR	A person with thyroidcancer has a/an chemotherapy treatment because/since/as chemotherapy is seldom helpful for most types of thyroid cancer, but fortunately it is not needed in most cases. It is often combined with external beam radiation therapy for anaplastic thyroid cancer and is sometimes used for other advanced cancers that no longer respond to other treatments.
T5-L	A person with thyroidcancer has a chemotherapy treatment because chemotherapy destroys the cells that cause the cancer.
T5-B	A person with thyroidcancer has a chemotherapy treatment because chemotherapy reduces the risk of thyroid cancer.
BART-L	A person with thyroidcancer has a chemotherapy treatment because chemotherapy can slow down the growth of thyroid cancer
BART-B	A person with thyroidcancer has a chemotherapy treatment because chemotherapy can help reduce the risk of thyroid cancer
Input	A person with depression has a/an medical conditions risk factor because/since/as {explanation}
Human	A person with depression has a medical conditions risk factor because the pain and stress that come with certain conditions, including cancer, stroke, or heart disease, can take a toll on a person's mental state resulting in depression.
RETR	A person with depression has a/an medical conditions risk factor because/since/as more commonly studied were the connections between depression and cardiovascular disease, metabolic syndrome, biochemical alterations, diabetes, dementia, cognitive impairment, Alzheimer's disease, somatization and chronic pain, asthma, arthritis, and hyperlipidemia.
T5-L	A person with depression has a medical conditions risk factor because depression can be caused by a number of medical conditions, such as diabetes, heart disease, high blood pressure, high cholesterol and high blood sugar.
T5-B	A person with depression has a medical conditions risk factor because medical conditions such as diabetes, heart disease, diabetes, AIDS, and other conditions that cause depression, such as cancer, are linked to depression.
BART-L	A person with depression has a medical conditions risk factor because depression can be caused by a variety of medical conditions.
BART-B	A person with depression has a medical conditions risk factor because depression can be a cause of depression.
Input	A person with depression has a/an grief risk factor because/since/as {explanation}
Human	A person with depression has a grief risk factor because grief associated with loss of dear one greatly increases the risk of psychiatric complications such as depression
RETR	A person with depression has a/an grief risk factor because/since/as risk Factors · Genetics: A history of depression in your family may make it more likely for you to get it. · Death or loss: Sadness and grief are ...
T5-L	A person with depression has a grief risk factor because people who have experienced a loss of a loved one are more likely to develop depression.
T5-B	A person with depression has a grief risk factor because grief is the most common cause of depression.
BART-L	A person with depression has a grief risk factor because grief can cause depression
BART-B	A person with depression has a grief risk factor because grief is associated with depression.

Table 12: More qual examples for CHARDAT$_{test}$: human, RETR, T5-large (L), T5-base (B), BART-large (L), BART-base (B)