Knowledge-Infused Prompting Improves Clinical Text Generation with Large Language Models

Ran Xu[♡], **Hejie Cui**[♡], **Yue Yu**[♠], **Xuan Kan**[♡], **Wenqi Shi**[♠], **Yuchen Zhuang**[♠], **Wei Jin**[♡], **Joyce C. Ho**[♡], **Carl Yang**[♡]
♡ Emory University ♠ Georgia Institute of Technology
{ran.xu,hejie.cui,xuan.kan,wei.jin,joyce.c.ho,j.carlyang}@emory.edu
{yueyu,wshi83,yczhuang}@gatech.edu

Abstract

Clinical natural language processing requires methods that can address domain-specific challenges, such as complex medical terminology and clinical contexts. Recently, large language models (LLMs) have shown promise in this domain. Yet, their direct deployment can lead to privacy issues and are constrained by resources. To address this challenge, we propose CLINGEN, which infuses knowledge into synthetic clinical text generation using LLMs for clinical NLP tasks. Our model involves clinical knowledge extraction and context-informed LLM prompting. Both clinical topics and writing styles are drawn from external domain-specific knowledge graphs and LLMs to guide data generation. Extensive studies across 7 clinical NLP tasks and 16 datasets reveal that CLINGEN consistently enhances performance across various tasks, effectively aligning the distribution of real datasets and enriching the diversity of generated training instances.

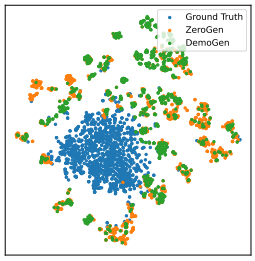
1 Introduction

Clinical Natural Language Processing (NLP) emerges as a distinct subfield including the extraction, analysis, and interpretation of medical data from unstructured clinical text [63]. Despite its significance, unique challenges evolve for methodology development in clinical NLP. For example, clinical texts are often dense with abbreviations and specialized medical terminologies that can be perplexing to standard NLP models [25]. Fortunately, recent advances in Large Language Models (LLMs) [8, 43, 42, 41] provide a promising way to resolve these issues, as they contain billions of parameters and have been pretrained on massive corpora, thus inherently capture a significant amount of clinical knowledge [2, 40, 15, 62, 53, 54, 32, 30].

Despite the strong capacity of general LLMs, directly applying them to infer on clinical text is often undesired in practice. Firstly, these LLMs often have billions of parameters, leading to *increased infrastructure costs* and *long inference time*. Furthermore, the sensitive patient information contained in the clinical text naturally raises *privacy concerns* [36, 20]. To effectively combat these challenges, generating synthetic training data using LLMs stands out as a promising solution: It leverages the capability of LLMs in a way that is both resource-efficient and privacy-centric.

Synthetic data generation with foundation models is a popular research domain in general machine learning [6, 34, 66, 67]. However, when considering producing high-quality data that conforms to the distribution of the original dataset, simply adapting LLMs to generate clinical data presents unique challenges. To assess the quality of data generated by current methods, we carry out an evaluation centered on distribution and diversity, detailed in Section 2. We further examine the clinically-related entity quantities and frequencies in synthetic data, where a notable decline is observed when contrasting synthetic data with ground truth data. Till now, a unified principle to better adapt LLMs for generating synthetic text for facilitating clinical downstream applications is still missing.

NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.



(a) t-SNE plot with Sentence-BERT [49] Embeddings.

	ZeroGen	DemoGen	Ground Truth
Entail	**Sentence A:** Can drinking alcohol increase the risk of liver disease? **Sentence B:** Does alcohol consumption contribute to liver disease risk?	**Sentence A:** What are the side effects of chemotherapy? **Sentence B:** What are the possible adverse effects of chemotherapy?	**Sentence A:** My 3yrs old boy found my bleach at the laundry and I suspect he swallowed a bit of it. How do I treat this pls. **Sentence B:** What the Doc will do if a child swallows bleach?
Not Entail	**Sentence A:** What are the side effects of metformin? **Sentence B:** Can I take ibuprofen for a headache?	**Sentence A:** What are the common symptoms of influenza? **Sentence B:** Can I take ibuprofen to manage my headache?	**Sentence A:** I have exercise induced asthma. Would any of these non drug devises be suitable please? **Sentence B:** Are there any treatments or cures for albinism?

(b) Case study of generated examples

Figure 1: Preliminary Studies. (a) is from BC5CDR-Disease and (b) is from MEDIQA-RQE.

Motivated by the above analysis, we propose CLINGEN, a *clinical knowledge-infused* generic framework for high-quality clinical text generation in few-shot scenarios. Our ultimate goal is to narrow the gap between synthetic and ground-truth data and encourage the topic diversity of the generated text. Towards this end, we propose a strategy to utilize clinical knowledge extraction to contextualize the prompts. This includes generating clinical topics from both KGs and LLMs and deriving writing style suggestions from LLMs. By doing this, CLINGEN integrates both non-parametric insights from external clinical knowledge graphs with the intrinsic parametric knowledge encoded in large language models. It is worth noting that, CLINGEN only rely on minimal additional human efforts, and can be readily applied to a wide array of core tasks in clinical NLP.

We conduct an exhaustive evaluation of synthetic clinical data generation **across 7 clinical NLP tasks and 16 datasets**. Empirical findings demonstrate that CLINGEN not only aligns more closely with the distribution of the original data but also amplifies the diversity of the generated training samples. The empirical performance gains are consistent across various tasks with different LLMs and classifiers (8.98% for PubMedBERT$_{\texttt{Base}}$ and 7.27% for PubMedBERT$_{\texttt{Large}}$).

2 Preliminary Study

2.1 Problem Setup

In this paper, we study the synthetic data generation problem in the few-shot setting. The input consists of a training set $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{K}$, where each (x_i, y_i) pair represents an input text and its corresponding label for the i-th example. K denotes the total number of training samples, which is intentionally kept at a very small value (5-shot per label). The primary objective is to harness the capabilities of an LLM $\mathcal{M}$ to generate a synthetic dataset, denoted as $\mathcal{D}_{\text{syn}} = \{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{N}$, where N is the number of generated samples ($N \gg K$). For each downstream task, we fine-tune an additional pre-trained classifier $\mathcal{C}_\theta$ parameterized by θ on the synthetic dataset $\mathcal{D}_{\text{syn}}$.

2.2 Limitations of Existing Synthetic Data Generation Methods

Here, we take a closer look at the synthetic text data generated by two representative approaches: ZeroGen [66] which directly prompts LLMs for data generation, and DemoGen [68, 35], which augments the prompt with few-shot demonstrations. We observe that these methods often introduce distribution shifts and exhibit limited diversity, which can be suboptimal for improving downstream performance. The illustration is as follows, and we include additional figures in Appendix C.

Distribution Shift. An inherent challenge when adapting LLMs to specific domains for text generation is the issue of *distribution shift*. In Figure 1(a), we visualize the embeddings of both the ground truth and synthetic datasets generated via two representative methods. Overall, these methods use generic prompts (see Appendix G.3 for details) with minimal domain-specific constraints. This limitation remains evident even when incorporating few-shot demonstrations into the process, with a notable disparity between the embeddings of the ground truth data and synthetic data.

Limited Diversity. Clinical datasets in real-world scenarios possess valuable knowledge that can be challenging to replicate within synthetically generated data by AI models. To explicitly illustrate this point, we present a case study in Figure 1(b). The comparison reveals that the samples generated by ZeroGen and DemoGen tend to be more straightforward, lacking the *sufficient details* present in

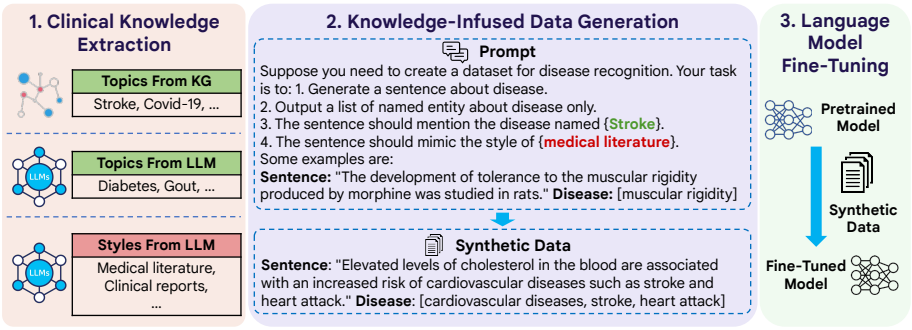


Figure 2: The overview of CLINGEN.

the ground truth data. Furthermore, the generated samples have a more uniform style and structure, while the ground truth encompasses various writing styles, including urgent and informal inquiries.

3 Clinical Knowledge Infused Data Generation

We introduce our novel framework, CLINGEN, a prior knowledge-informed clinical data generation. The overview of CLINGEN is shown in Figure 2. This two-step methodology harnesses the emergent capabilities of LLMs and external knowledge from KGs to facilitate the synthesis of clinical data.

3.1 Clinical knowledge extraction

Contrary to previous studies [66, 34, 35] which employ generic queries to prompt LLMs for text generation, CLINGEN aims to extract rich clinically relevant knowledge from parametric (*e.g.* LLMs) or nonparametric sources (*e.g.* knowledge graphs) and tailor it for prompting LLM to clinical NLP tasks. To realize this objective, our modeling contains two dimensions including *clinical topics* and *writing styles*, which are integrated into the original prompts to infuse domain-specific knowledge. By dynamically composing different topics and writing styles together, CLINGEN can provide a diverse suite of prompts, resulting in a wider spectrum of text produced from LLM.

3.1.1 Clinical Topics Generation

We provide two choices to generate clinical topics – one is to sample related entities or relations from external KG, and the other is to query relevant knowledge from LLM.

Topics sampled from Non-Parametric KGs. Healthcare KGs offer a rich collection of medical concepts and their complex relationships, and have emerged as a promising tool for organizing medical knowledge in a structured way [13]. We employ the iBKH KG [55] due to its broad coverage over clinical entities. To illustrate, for the Disease Recognition task (NCBI) [14], we extract all medication nodes from the iBKH to bolster the pharmaceutical information. As another example, we retrieve links between drug and disease nodes for the chemical and disease relation extraction (CDR) task [60]. By integrating the information from the clinical KG into the generation process, the generated samples exhibit a high degree of contextual accuracy, diversity, and semantic richness.

Topics queried from Parametric Model (LLMs). LLMs provide an alternative method for acquiring domain knowledge, as they are pre-trained on extensive text corpora, including medical literature. Specifically, we aim to harness the rich clinical domain knowledge encoded in ChatGPT (`gpt-3.5-turbo-0301`) to augment the prompt. The incorporated prior knowledge from LLMs is focused on entity categories that hold significant relevance within clinical text datasets, including diseases, drugs, symptoms, and side effects. For each of these pivotal entity types, we prompt the LLMs by formulating inquiries, *e.g.*, "`Suppose you are a clinician and want to collect a set of <Entity Type>. Could you list 100 entities about <Entity Type>?`". These crafted conversational cues serve as effective prompts, aiding in the retrieval of clinically significant entities from the extensive domain knowledge within LLMs.

3.1.2 Writing Styles Suggestion

Styles suggested by LLMs. To address the limitations mentioned in Sec 2.2 and introduce a diverse range of writing styles into the generated samples, we leverage the powerful LLM again by suggesting candidate writing styles for each task. Specifically, we incorporate task names into our prompts

(*e.g.*, disease entity recognition) and integrate few-shot demonstrations. We then engage ChatGPT in suggesting several potential sources, speakers, or authors of the sentences. See Appendix G.1 for detailed prompt. Responses such as "`medical literature`" or "`patient-doctor dialogues`" are augmented into the prompts to imitate the writing styles found in real datasets.

3.2 Knowledge-infused Synthetic Data Generation

With the generated entities as well as styles, the key challenge becomes how to leverage them to extract rich clinical information from the LLM for improving synthetic data quality. Directly putting all the elements to enrich the prompt is often infeasible due to the massive size of entities. To balance informativeness as well as diversity, we propose a knowledge-infused strategy, where the collected clinical topics and writing styles serve as the base unit. In each step, we randomly sample a topic and a writing style from the candidate set to augment the prompts. For instance, for the Disease Recognition (NCBI) task, consider a clinical entity like "`stroke`" . We enrich the prompt query for LLM by appending "`generate a sentence about stroke`" as a generation guidance. For a comprehensive view of the prompt formats across various tasks, please refer to Appendix G. Despite its simplicity, this knowledge-infused strategy ensures that the clinical context is incorporated into the prompts while encouraging prompt diversity (via composing different entities and writing styles), thereby enhancing the quality and clinical relevance of the generated synthetic data.

3.3 Language model fine-tuning

After generating synthetic data $\mathcal{D}_{\text{syn}}$ through LLMs, we fine-tune a pre-trained classifier $\mathcal{C}_\theta$ to each downstream task. Following [35], we first fine-tune $\mathcal{C}_\theta$ on $\mathcal{D}_{\text{train}}$ with standard supervised training objectives (denoted as $\ell(\cdot)$), then on the synthetic data $\mathcal{D}_{\text{syn}}$. We strictly follow a standard fine-tuning process and avoid using any extra techniques: (1) for standard classification tasks, $\ell(\cdot)$ is the cross-entropy loss; (2) for multi-label classification tasks, $\ell(\cdot)$ is the binary cross-entropy loss; (3) for token-level classification tasks, we stack an additional linear layer as the classification head and $\ell(\cdot)$ is the token-level cross-entropy loss. The design of *advanced learning objectives* as well as *data mixing strategies*, while important, are orthogonal to the scope of this paper.

4 Empirical Evaluation

4.1 Experiment Setup

We conduct experiments in the few-shot settings with 5 examples available for each class. We employ ChatGPT [42] (`gpt-3.5-turbo-0301`) as the generator for both CLINGEN and baselines for a fair comparison. The pre-trained PubMedBERT [18] is then applied to fine-tune the generated synthetic data for evaluation. See Appendix D for implementation details.

Datasets and Tasks. In our exploration of few-shot synthetic data generation, we undertake a comprehensive evaluation of **16 datasets** across a diverse array of tasks typically encountered in clinical NLP benchmarks [45, 17]. Please see Appendix E for detailed dataset descriptions.

Baselines. We compare CLINGEN with **7 baselines** in total, including 4 data augmentation methods and 3 LLM-based data generation techniques. See Appendix F for details.

4.2 Experimental Results

Model Performance with the Synthetic Data Table 1 summarizes the experimental results on different datasets. We also conduct supervised learning on the original training data and the extracted few-shot examples, denoted as "Supervised-Full" and "Supervised-Few", respectively. Due to space limits, we report the average performance over all datasets for each task, but provide the detailed results for each dataset in Tables 6, 7, 8 in Appendix H. We have the following findings:

◇ Our proposed approach, CLINGEN, consistently outperforms the baselines across all tasks. The average performance gain over all *main* metrics is 8.98% at `Base` scale and 7.27% at `Large` scale. In addition, methods utilizing LLMs have better performance than traditional data augmentation techniques, illustrating the capacity of LLMs to extract valuable information from limited examples.

◇ In *token classification tasks*, CLINGEN performs better with KG compared to LLM. This improvement stems from the strong alignment between the task's target and the generated domain knowledge, where the extracted topics serve as direct labels for these datasets. The *single-sentence*

Task	Text Class	RE	NLI	Fact Verification		STS	NER	MedAttr		
	F1	F1	Acc	Acc	F1	Acc	F1	P	R	F1
PubMedBERT$_{\texttt{Base}}$										
Supervised-Full	77.01	77.34	79.20	67.58	65.49	75.70	89.67	—	—	—
Supervised-Few	18.61	43.89	44.64	29.43	27.10	55.70	39.41	38.11	43.82	40.77
DA-Word Sub	40.74	38.14	55.08	28.86	25.83	54.40	44.30	40.25	47.65	43.64
DA-Back Trans	47.24	—	54.30	32.15	28.04	55.80	—	—	—	—
DA-Mixup	45.09	43.37	53.52	32.78	29.12	58.20	42.20	42.37	48.96	45.43
DA-Transformer	41.02	47.56	55.71	35.32	31.77	58.80	44.75	37.82	44.28	40.80
ZeroGen	59.02	63.84	55.96	35.30	32.50	68.35	56.97	52.80	49.53	51.11
DemoGen	64.09	67.46	59.80	40.30	35.95	70.85	60.16	58.15	56.84	57.49
ProGen	65.16	67.23	59.57	37.71	34.54	69.30	60.49	57.76	58.57	58.16
ClinGen w/ KG	67.15	69.01	64.89	43.83	39.43	72.17	**64.26**	**71.75**	65.20	**68.32**
ClinGen w/ LLM	**67.82**	**70.06**	**67.24**	**46.50**	**41.46**	**73.31**	63.17	68.19	**66.79**	67.48
Performance Gain	4.08%	3.85%	12.44%	15.38%	15.33%	3.47%	6.23%	—	—	17.47%
PubMedBERT$_{\texttt{Large}}$										
Supervised-Full	80.06	79.64	82.65	72.97	69.23	78.80	90.15	—	—	—
Supervised-Few	17.86	52.68	50.00	40.90	30.50	59.73	42.84	41.30	45.02	43.08
DA-Word Sub	43.99	44.35	57.66	35.51	31.95	55.30	46.67	46.77	43.52	45.09
DA-Back Trans	50.98	—	58.39	34.12	31.36	56.40	—	—	—	—
DA-Mixup	46.74	50.97	57.35	34.01	31.10	58.50	46.69	41.25	52.09	46.04
DA-Transformer	44.41	46.12	58.94	35.09	30.95	58.10	46.94	43.36	45.78	44.54
ZeroGen	61.51	65.18	63.47	41.12	36.10	72.69	57.79	54.04	51.40	52.69
DemoGen	64.97	68.65	64.58	42.61	38.69	74.37	61.43	62.67	61.02	61.83
ProGen	65.01	69.23	63.32	42.79	38.63	74.89	62.47	57.21	63.70	60.28
ClinGen w/ KG	66.76	71.47	**70.90**	48.62	42.45	75.82	**65.48**	70.96	**69.66**	**70.30**
ClinGen w/ LLM	**67.61**	**72.81**	70.50	**49.51**	**43.72**	**76.21**	65.36	**71.61**	66.86	69.15
Performance Gain	4.00%	5.17%	9.79%	15.70%	13.00%	3.47%	1.76%	—	—	13.70%

Table 1: Experimental results aggregated by tasks. **Bold** and underline indicate the best and second best results for each dataset, respectively. †: The models can only be applied to NER tasks, and the number is reported from the original paper. $*$: Since the two † models only report results on two NER datasets, we report an average performance on those two datasets for a fair comparison.

	HOC	GAD			ChemProt	MEDIQA-RQE	PUBHEALTH		NCBI-Disease			CASI		
	F1	P	R	F1	F1	ACC	ACC	F1	P	R	F1	P	R	F1
GPT-3.5 Inference	68.76	84.21	**97.46**	90.35	49.42	74.31	**69.50**	**52.47**	46.62	52.31	49.30	48.82	**74.75**	59.07
ClinGen w/ KG	77.71	94.30	89.09	**91.62**	60.12	**79.92**	50.20	41.26	**62.46**	**64.08**	**63.26**	70.96	69.66	**70.30**
ClinGen w/ LLM	**78.14**	**95.08**	86.14	90.39	**63.05**	77.36	52.96	43.31	61.12	60.16	60.64	**71.61**	66.86	69.15

Table 2: Comparison between prompting GPT-3.5 for inference and ClinGen at `Large` scale.

and *sentence-pair tasks*, on the other hand, display an advantage for the LLM-based knowledge extraction due to two reasons: (1) these tasks prioritize understanding entire sentences over specific terminologies, and some specialized terms might even impede LLM comprehension. (2) KGs may not always contain the required information. For example, in an RE task involving chemicals and proteins, some types of relations are absent from the KG, thus the performance gain is rather limited.

Comparison with few-shot inference via prompting GPT-3.5. As we employ GPT-3.5 as our backbone model for generating synthetic data, we also evaluate its ability for direct inference from the original training sets using few-shot in-context learning. Due to budget limits, we only run experiments on datasets with few testing samples for each task. As presented in Table 2, ClinGen at PubMedBERT$_{\texttt{Large}}$ scale achieves better results on 5 out of 6 datasets than GPT-3.5 few-shot learning, which uses $\sim 530\times$ more parameters. One exception is for PUBHEALTH, as it requires complex reasoning abilities that PubMedBERT$_{\texttt{Large}}$ may not fully possess. Overall, ClinGen offers cost-effective and time-efficient advantages, as subsequent prediction relying on a moderate-sized pretrained language model is much more efficient. Besides, the continued use of GPT-3.5 for inference on new testing data incurs ongoing time and financial costs, while our model requires zero additional costs for querying APIs. The price information is exhibited in Appendix K.

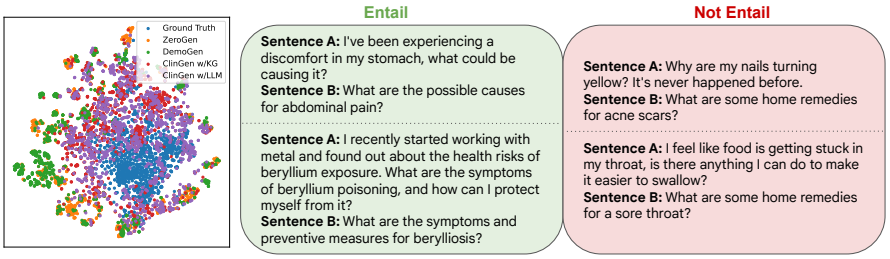
Ablation Study. We inspect different components of ClinGen in Table 3. It is observed that both Topics Extraction and Style Suggestion contribute to model performance. Different from the other datasets, MEDIQA-RQE shows more performance gain incorporating writing style than topics. It is because NLI tasks focus on capturing the relationships between two sentences while incorporating additional knowledge entities does not directly help the model improve the reasoning ability.

	HOC		CDR		MEDIQA-RQE		NCBI-Disease	
	w/ KG	w/ LLM	w/ KG	w/ LLM	w/ KG	w/ LLM	w/ KG	w/ LLM
ClinGen	**76.28**	**76.42**	**61.74**	**63.34**	**74.85**	**72.40**	**59.46**	**55.95**
w/o Styles	73.25	74.40	59.10	60.15	67.21	66.50	57.97	54.70
w/o Topics	70.86		58.51		69.86		55.09	

Table 3: Ablation studies on topic extraction and style suggestion at `Base` scale.

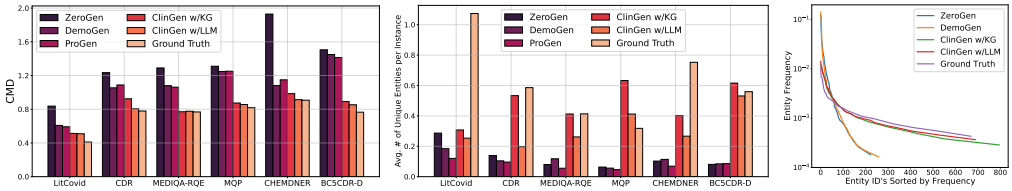
	HOC	CDR	MEDIQA-RQE	NCBI-Disease
ZeroGen	0.512	0.469	0.277	0.528
DemoGen	0.463	0.377	0.289	0.281
ProGen	0.481	0.321	0.290	0.357
ClinGen w/ KG	0.440	**0.291**	**0.243**	0.180
ClinGen w/ LLM	**0.432**	0.338	0.255	**0.155**
Ground truth	0.265	0.268	0.164	0.262

Table 4: Average Pairwise Similarity.



(a) t-SNE plot (b) Case study of generated examples

Figure 3: Data distribution and diversity measures on CLINGEN. (a) is from BC5CDR-Disease and (b) is from MEDIQA-RQE using CLINGEN with LLM.



(a) CMD (b) Entity Coverage (c) Entity Frequency

Figure 4: Data distribution and diversity measures on CLINGEN. (c) is from BC5CDR-Disease.

4.3 Quality Analysis of the Synthetic Data

Data Distribution Measures. In this section, we present the data distribution and diversity measurement of the synthetic dataset generated by CLINGEN. Figure 3(a) shows the t-SNE plot of data generated by CLINGEN and baselines compared with the ground truth. This visualization clearly demonstrates that CLINGEN exhibits a greater overlap with the ground truth, indicating a similar distribution as the original dataset. In addition, as depicted in Figure 4(a), the embedding of CLINGEN aligns more closely with the ground truth distribution than other baselines across all six datasets, further justifying the efficacy of CLINGEN for mitigating the distribution shift issue.

Diversity Measures. Table 4 calculates the average cosine similarity for sample pairs using SentenceBERT embeddings. Compared to baselines, the dataset generated with CLINGEN exhibits lower cosine similarity and the average similarity is close to that of the ground truth training data, which shows CLINGEN could render more diverse data. Moreover, Figure 4(b) highlights the ability of CLINGEN to cover a broader range of entities in comparison to the baselines. We find that CLINGEN with KG captures a larger variety of entities than CLINGEN with LLM, because KG tends to cover more extensive knowledge, including relatively uncommon information that may not be present in LLMs. Figure 4(c) reflects that the entity frequency distribution of CLINGEN is more in line with the ground truth, having a relatively balanced distribution among all entities. This ensures that CLINGEN generates synthetic data with a wide range of diverse topics.

Case Study. In Figure 3(b), we present a case study of examples generated by CLINGEN with LLM on MEDIQA-RQE dataset, which consists of consumer health queries. The sentences generated by CLINGEN include more extensive contextual information compared with the baseline as shown in Figure 1(b). These sentences closely resemble the queries people might pose in real-life scenarios.

5 Conclusion

We introduce CLINGEN, a new framework that leverages clinical knowledge from non-parametric KGs and parametric LLMs. This knowledge empowers data generation by utilizing clinical topic knowledge and real-world writing styles in domain-specific prompts. Our extensive evaluations across 7 clinical NLP tasks and 16 datasets, consistently show that CLINGEN improves task performance, aligns closely with real data, and enhances data diversity. We expect this approach can be seamlessly incorporated into a broad suite of clinical text tasks to advance clinical NLP research.

References

[1] Asma Ben Abacha and Dina Demner-Fushman. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association, 2016.

[2] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[3] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 17–23 Jul 2022.

[4] Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440, 2015.

[5] Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy, August 2019. Association for Computational Linguistics.

[6] Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1), February 2015.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33, 2020.

[9] Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, 2020.

[10] Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. Overview of the biocreative vii litcovid track: multi-label topic classification for covid-19 literature annotation. In *Proceedings of the BioCreative challenge evaluation workshop*, 2021.

[11] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR, 2021.

[12] John Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] Hejie Cui, Jiaying Lu, Shiyu Wang, Ran Xu, Wenjing Ma, Shaojun Yu, Yue Yu, Xuan Kan, Tianfan Fu, Chen Ling, et al. A survey on knowledge graphs for healthcare: Resources, application progress, and promise. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[14] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

[15] Lehman; Eric, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models? In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209, pages 578–597. PMLR, 2023.

[16] Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*, 2017.

[17] Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, et al. Bigbio: A framework for data-centric biomedical natural language processing. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[18] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[19] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823, 2021.

[20] Geoff Keeling. Algorithmic bias, generalist models, and clinical medicine. *arXiv preprint arXiv:2305.04008*, 2023.

[21] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[22] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7740–7754, Online, November 2020. Association for Computational Linguistics.

[23] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Batista-Navarro, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1):S2, Jan 2015.

[24] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China, December 2020.

[25] Peter Lee, Carey Goldberg, and Isaac Kohane. *The AI Revolution in Medicine: GPT-4 and Beyond*. Pearson Education, Limited, 2023.

[26] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016, 2016.

[27] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[28] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1054–1064, 2020.

[29] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, 2022.

[30] Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, et al. Radiology-gpt: A large language model for radiology. *arXiv preprint arXiv:2306.08666*, 2023.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23:bbac409, 2022.

[33] Clara H McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. Effective transfer learning for identifying similar questions: matching user questions to covid-19 faqs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3458–3465, 2020.

[34] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[35] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR, 2023.

[36] Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ Digital Medicine*, 6(1):120, 2023.

[37] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, 2013.

[38] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

[39] Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307, 2014.

[40] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[41] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[42] OpenAI. Introducing chatgpt, 2023.

[43] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[45] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.

[46] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, 2021.

[47] Zhaozhi Qian, Thomas Callender, Bogdan Cebere, Sam M Janes, Neal Navani, and Mihaela van der Schaar. Synthetic data for privacy-preserving clinical risk prediction. *medRxiv*, 2023.

[48] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

[49] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992, 2019.

[50] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[51] Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[52] Chaitanya Shivade. Mednli — a natural language inference dataset for the clinical domain, 2017.

[53] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 2023.

[54] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[55] Chang Su, Yu Hou, Manqi Zhou, Suraj Rajendran, Jacqueline RMA Maasch, Zehra Abedi, Haotan Zhang, Zilong Bai, Anthony Cuturrufo, Winston Guo, et al. Biomedical discovery through the integrative biomedical knowledge hub (ibkh). *Iscience*, 26(4), 2023.

[56] Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S Roque, Irene Kouskoumvekaki, Alina Bora, et al. Chemprot: a disease chemical biology database. *Nucleic acids research*, 39:D367–D372, 2010.

[57] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023.

[58] Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize extremely high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *arXiv preprint arXiv:2304.02169*, 2023.

[59] Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, 19:1–13, 2019.

[60] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016:baw032, 2016.

[61] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[62] Cliff Wong, Sheng Zheng, Yu Gu, Christine Moung, Jacob Abel, Naoto Usuyama, Roshanthi Weerasinghe, Brian Piening, Tristan Naumann, Carlo Bifulco, et al. Scaling clinical trial matching using large language models: A case study in oncology. *arXiv preprint arXiv:2308.02180*, 2023.

[63] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of clinical foundation models: A survey of large language models and foundation models for emrs. *arXiv preprint arXiv:2303.12961*, 2023.

[64] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[65] Ran Xu, Yue Yu, Joyce Ho, and Carl Yang. Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2501–2505, 2023.

[66] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, 2022.

[67] Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Progen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, 2022.

[68] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[69] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895*, 2023.

[70] Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022.

[71] Rongzhi Zhang, Yue Yu, and Chao Zhang. SeqMix: Augmenting active sequence labeling via sequence mixup. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8566–8579, Online, November 2020. Association for Computational Linguistics.

[72] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[73] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A Additional Related Work

Generating additional training data enables a more precise analysis of medical text, and has gained more attention in the past years. Earlier research has employed data augmentation techniques to generate similar samples to existing instances with word substitution [19, 50], back translation [64], pretrained transformers [24, 73] for enhancing model generalization. But they often yield rigid transformations and the quality of the augmented text cannot be always guaranteed. Another line of work focuses on leveraging external knowledge to create weak labels [48, 16, 59, 65][1]. These methods typically require domain expertise and additional task-specific corpora, which can be resource-intensive to obtain for low-resource clinical tasks. Moreover, designing effective rules can be challenging due to the reliance on domain-specific knowledge [70].

The emergence of generative AI has presented new possibilities, and some studies attempt to use generative models for generating synthetic training data on both general domains [34, 35, 66, 69, 12], and clinical domains [11, 57, 58, 47]. However, these methods often use generic and simple prompts that may not fully capture domain-specific knowledge, thus potentially limiting the quality of the generated data. [29, 12] employ interactive learning to generate additional instances to refine the existing dataset, at the cost of additional human efforts. One recent study [57] explores synthetic data generation for clinical NLP. Nevertheless, their proposed approach relies on a *much larger training set* to generate candidate entities, which disregards the practical low-resource setting [46]. Furthermore, their study is limited to a narrow range of tasks (2 tasks and 4 datasets only), lacking breadth in terms of exploring a diverse set of clinical NLP applications.

B Limitation, Future Works and Ethics Issues

In this work, we propose CLINGEN to better harness the LLM for synthetic text data generation. Despite the strong performance of CLINGEN on 16 clinical NLP tasks, we mainly verify their efficacy from their empirical performance, sample diversity, and distribution gaps. However, there still exist gaps between the performance of the model $\mathcal{C}_\theta$ fine-tuned using our generated synthetic data and ground-truth data. To further improve CLINGEN, there are several avenues of future works:

Using Clinical LLMs as Data Generator: Our method CLINGEN relies on an LLM with instruction following ability. We mainly evaluate CLINGEN using GPT-family models as the LLM. Recently, there are many LLMs that have been fine-tuned on additional clinical contexts as well as instructions (e.g. Med-PALM[2]), and achieved superior performance on challenging clinical NLP benchmarks. However, they are not open-sourced, thus we cannot run them in our experiments. An interesting future work could be how to leverage these Clinical LLMs as Data Generator to further boost the performance.

Data Evaluation: In this work, we consider the distribution gap and sample diversity as our optimization objective. However, there might be many other aspects for synthetic quality estimation [3]. We need more tools to capture, analyze, and improve this new aspect of data-centric AI.

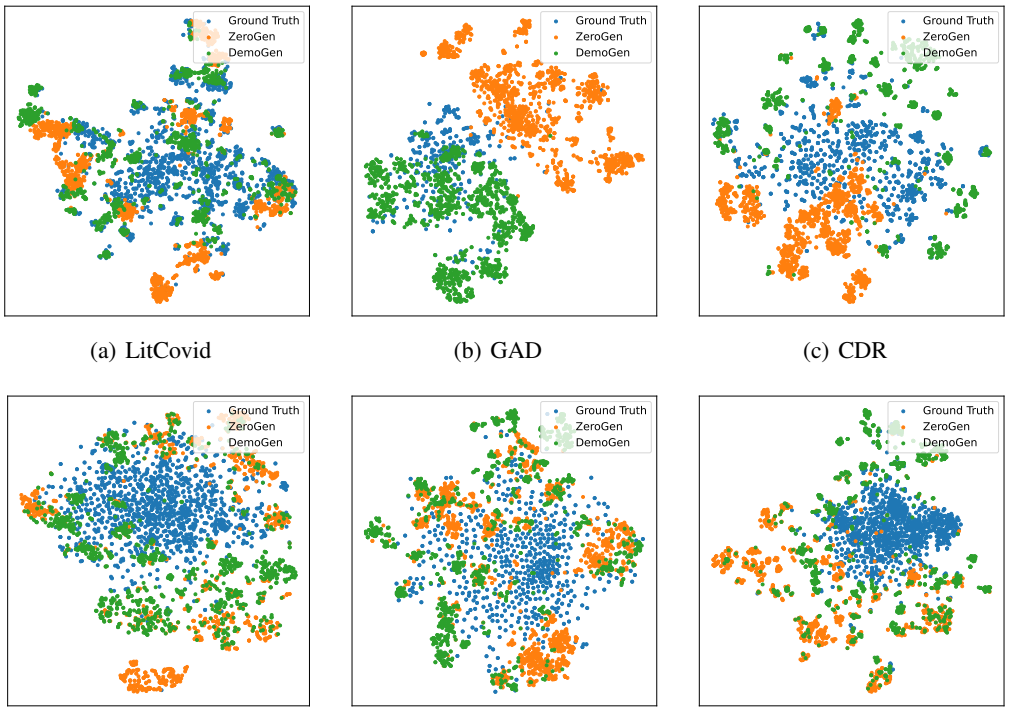
Factuality: One issue with LLM-based synthetic data generation is the phenomenon of *hallucination*, wherein the model generates information that does not ground in reality [72]. This can lead to the propagation of misinformation, which may have negative impacts on the clinical domain. It is crucial to cross-verify the generated text with a reliable knowledge base or dataset. Furthermore, incorporating an additional layer of human review can also help in mitigating hallucinations and ensuring the faithfulness of LLM-generated synthetic outputs.

C Additional Preliminary Studies

We present additional preliminary studies of the t-SNE plots in Figure 5 and the regularized entity frequencies in Figure 6. These results further justify the distribution shift issue mentioned in section 2.2, demonstrating that the limited diversity as well as the distribution shift issue generally exists for a broad range of clinical NLP tasks.

[1] Some works also name it as 'distant supervision' [38, 37, 28].

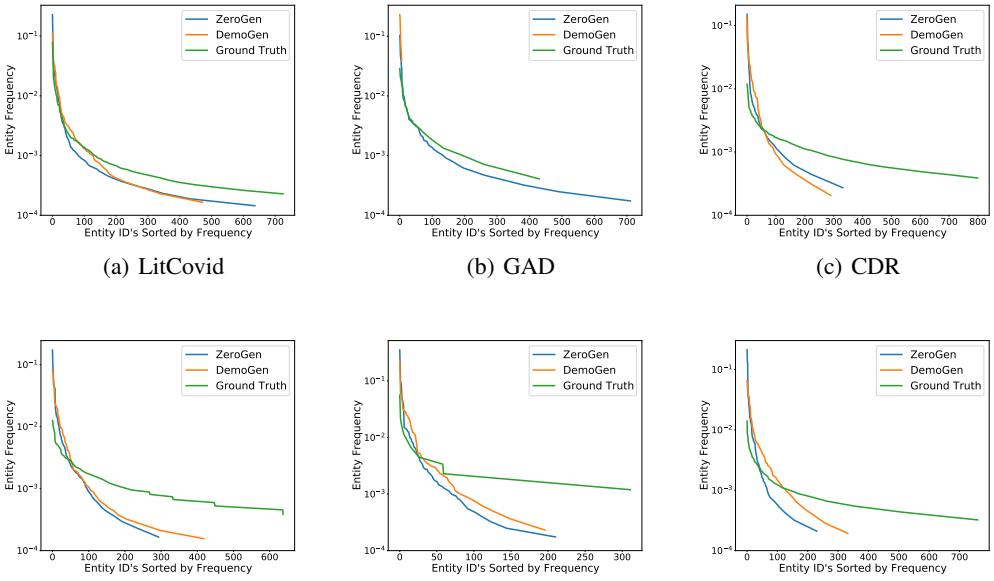
[2] https://sites.research.google/med-palm/



(a) LitCovid (b) GAD (c) CDR

(d) MEDIQA-RQE (e) MQP (f) CHEMDNER

Figure 5: The t-SNE plots of datasets generated by ZeroGen and DemoGen compared with the ground truth.



(a) LitCovid (b) GAD (c) CDR

(d) MEDIQA-RQE (e) MQP (f) CHEMDNER

Figure 6: The regularized entity frequencies of datasets generated by ZeroGen and DemoGen compared with the ground truth in log scale.

D Implementation Details

For implementation, we use PyTorch [44] and HuggingFace [61]. For each dataset, we randomly sample 5 examples from each class to provide few-shot demonstrations and keep a validation set of the same size. In the experiments, We generate 5000 synthetic training data for both CLINGEN and the baselines and report the average performance over 3 random seeds for all the results.

During the data generation process when we call the ChatGPT APIs [42], we set the parameter $\text{top\_p} = 1.0$ and temperature $t = 1.0$ to balance between the quality of the generated text as well as diversity [12, 69][3]. With the generated synthetic dataset, we follow the common few-shot learning setting [46] to train all the models for 6 epochs and use the model with the best performance on the validation set for evaluation.

During the PubMedBERT fine-tuning, we adopt AdamW [31] for optimization with a linear warmup of the first 5% steps and linear learning rate decay. The learning rate is set to 2e-5 for `Base` and 4e-5 for `Large`, and the maximum number of tokens per sequence is 256.

E Dataset Description

Corpus	Tasks	#Class	#Train/#Test	Metrics
Single-Sentence Tasks				
LitCovid [10]	Text Classification	7	24960/6238	**F1**
HOC [4]	Text Classification	10	3091/898	**F1**
GAD [7]	Relation Extraction (RE)	1	8431/2522	P, R, **F1**
CDR [60]	Relation Extraction (RE)	1	8431/2522	P, R, **F1**
ChemProt [56]	Relation Extraction (RE)	5	8793/1087	**F1**
Sentence-Pair Tasks				
MedNLI* [52]	Natural Language Inference (NLI)	3	11232/1422	**Acc**
MEDIQA-NLI† [5]	Natural Language Inference (NLI)	3	-/405	**Acc**
MEDIQA-RQE [1]	Natural Language Inference (NLI)	2	8588/302	**Acc**
PUBHEALTH [22]	Fact Verification	4	9804/1231	Acc, **F1**
HealthVer [51]	Fact Verification	3	10591/1824	Acc, **F1**
MQP [33]	Sentences Similarity (STS)	2	10/3033	**Acc**
Token Classification Tasks				
BC5CDR-Disease [26]	Named Entity Recognition (NER)	1	4882/5085	P, R, **F1**
BC5CDR-Chemical [26]	Named Entity Recognition (NER)	1	4882/5085	P, R, **F1**
NCBI-Disease [14]	Named Entity Recognition (NER)	1	5336/921	P, R, **F1**
CHEMDNER [23]	Named Entity Recognition (NER)	1	14522/12430	P, R, **F1**
CASI [2, 39]	Attribute Extraction	6	5/100	**F1**

Table 5: Dataset statistics. We do not count the non-entity/non-relation class for relation extraction and token classification tasks to align with existing works. P and R stand for Precision and Recall. Metrics in **bold** are considered as the main metrics. $*$ is not allowed to put into GPT and † does not provide training data, so we sample few-shot examples from the SciTail [21] instead.

The evaluation tasks and datasets are summarized in Table 5. Note that the number of training samples indicates the size of the *original* training set. Specifically, we consider the following datasets:

- **Single-Sentence Tasks**
  - Text Classification:
    - The *LitCovid* dataset [10] consists of COVID-19-related publications from PubMed. The task is to predict the topics of the sentences, including "Epidemic Forecasting", "Treatment", "Prevention", "Mechanism", "Case Report", "Transmission", and "Diagnosis".
    - The *HOC* dataset [4] also extracts sentences from PubMed articles, each annotated at the sentence level. The task is to predict the topics of the sentences, including "evading

[3] We do not further increase t, as previous analysis [12, 69] has shown that increasing t to larger value does not help with additional performance gain.

growth suppressors", "tumor promoting inflammation", "enabling replicative immortality", "cellular energetics", "resisting cell death", "activating invasion and metastasis", genomic instability and mutation", "inducing angiogenesis", "sustaining proliferative signaling", and "avoiding immune destruction".

- Relation Extraction:
  - The *GAD* [7] dataset is to predict whether there is a relation between the given disease and gene in the sentences.
  - The *CDR* [60] dataset is to predict whether the provided chemical can induce the disease in the sentences.
  - The *ChemProt* [56] dataset focuses on the chemical-protein relations, and the labels include "Upregulator", "Downregulator", "Agonist", "Antagonist", "Product_of" and "No relation".

- **Sentence-Pair Tasks**
  - Natural Language Inference (NLI):
    - The *MedNLI* [52] dateset consists of sentences pairs derived from MIMIC-III, where we predict the relations between the sentences. The labels include "entailment", "neutral" and "contradiction".
    - The *MEDIQA-NLI* [5] dataset comprises text-hypothesis pairs. Their relations include "entailment", "neutral" and "contradiction".
    - The *MEDIQA-RQE* [1] dataset contains NIH consumer health question pairs, and the task is to recognize if the first question can entail the second one.
  - Fact Verification:
    - The *PUBHEALTH* [22] encompasses claims paired with journalist-crafted explanations. The task is to predict the relations between the claim and evidence, including "Refute", "Unproven", "Support", and "Mixture".
    - The *HealthVer* [51] contains evidence-claim pairs from search engine snippets regarding COVID-19 questions. The relations between claims and evidences are chosen from "Refute", "Unproven", and "Support".
  - Sentence Similarity (STS):
    - the *MQP* [33] dataset comprises a collection of medical question pairs designed for identifying semantically similar questions. The task is to predict whether the two questions are equivalent or not.

- **Token Classification Tasks**
  - Named Entity Recognition (NER):
    - The *BC5CDR-Disease* [27] is to recognize diseases in the sentences.
    - The *BC5CDR-Chemical* [27] is to recognize chemicals in the sentences.
    - The *NCBI-Disease* [14] is to recognize diseases in the sentences.
    - The *CHEMDNER* [23] is to recognize chemicals in the sentences.
  - Attribute Extraction (MedAttr):
    - The *CASI* dataset [2, 39] aims to identify interventions including medication, dosage, route, freq, reason, duration

F Baseline Details

Data Augmentation Methods:

- **DA-Word Sub**: Word substitution, following Checklist [50], maintains a word list to generate new examples by filling in a template.
- **DA-Back Translation**: Following UDA [64], we employ back translation to augment the training data, including translating text from the target language to the source language and then back to the target language.
- **DA-Mixup** [9, 71]: We use the TMix version of MixText for data interpolation on the few-shot labeled dataset. For NER tasks, we employ SeqMix [71] that augments the queried samples by generating additional labeled sequences iteratively.

- **DA-Transformer (MELM)** [24, 73]: It introduces a conditional data augmentation technique that prepends class labels to text sequences for pre-trained transformer-based models. For NER tasks, it incorporates NER labels into the sentence context and predicts masked entity tokens by explicitly conditioning on their labels.

LLM-based Generation Methods.

- **ZeroGen** [66]: It generates a dataset using a carefully designed instruction and then trains a tiny task-specific model for zero-shot inference. We follow the prompting method mentioned in their original paper as implementation, which *does not consider* any style information as well as domain knowledge.
- **DemoGen** [35, 68]: It leverages LLMs to synthesize novel training data by feeding few-shot samples as demonstrations. Note that we focus on using the black-box LLM as the generator, thus we do not tune the LLM as [35].
- **ProGen** [67]: It divides the entire dataset generation process into multiple phases. In each phase, feedback from the previously generated dataset guides the generation towards higher quality. It is also in the few-shot setting.

We do not compare with [57] in the main experiments as it leverages entities extracted from the entire training set and violates the true few-shot learning setting.

G Prompt Format

G.1 The prompts for Writing Styles Suggestion with CLINGEN

Listing 1: Prompt Format for writing styles suggestion with CLINGEN.

```
Suppose you need to generate a synthetic clinical text dataset on
[task] tasks. Here are a few examples from the original training
set:
[demonstrations]
Please write three potential sources, speakers or authors of the
sentences.
```

`[task]`: The task names for each specific task. `[demonstrations]`: The few-shot demonstrations from the original training set.

G.2 The prompts for Data Generation with CLINGEN

In the following prompt format, `[topic]` and `[style]` are randomly sampled from the topics candidate set and styles candidate set we formulate in the knowledge extraction step, respectively.

Named entity recognition tasks:

Listing 2: Prompt Format for NER tasks with CLINGEN.

```
Suppose you need to create a dataset for [domain] recognition.
Your task is to:
1. generate a sentence about [domain],
2. output a list of named entity about [domain] only,
3. the sentence should mimic the style of [style],
4. the sentence should mention the [domain] named [topic].
```

`[domain]`: "disease" for BC5CDR-Disease and NCBI-Disease; "chemical" for BC5CDR-Chemical and CHEMDNER.

Medication attributes tasks:

Listing 3: Prompt Format for medication attributes tasks with CLINGEN.

```
Suppose you need to create a dataset for clinical attributes
recognition. Your task is to:
1. generate a sentence about clinical attributes, The Clinical
Attributes you need to extract include "Medication", "Dosage", "
Route", "Frequency", "Reason", "Duration". For each attribute
class, please return a list of attributes within the class that
occurs in the Sentence.
2. the sentence should mimic the style of [style],
3. the sentence should be relevant to [topic].
```

Text classification tasks:

Listing 4: Prompt Format for text classification tasks with CLINGEN.

```
Suppose you are a writer for [domain]. Your task is to:
1. give a synthetic [domain] about [class_name].
2. discuss about the subtopic of [topic] for [class_name] in the
[domain].
3. the sentence should mimic the style of [style].
```

`[domain]`: "COVID-19 Literature" for LitCovid and "Cancer Document" for HOC.

`[class_name]`: the label name for this generated sample.

Relation extraction tasks:

Listing 5: Prompt Format for relation extraction tasks with CLINGEN.

```
Suppose you need to generate synthetic data for the biomedical
[domain] task. Your task is to:
1. give a sentence about [class_name] relation between [entity0]
and [entity1]
2. the sentence should discuss the [entity0]: [topic0] and
[entity1]: [topic1] with the relation [label_desc].
3. the sentence should mimic the style of [style].
```

`[domain]`: "Disease Gene Relation" for GAD, "Chemical Disease Relation" for CDR, and "Chemical Protein Relation" for ChemProt.

`[entity0]` and `[entity1]`: "disease" and "gene" for GAD, "chemical" and "disease: for CDR, and "chemical" and "protein" for ChemProt.

`[class_name]`: the label name for this generated sample.

`[label_desc]`: the description of the selected label. For example, the label "upregulator" in ChemProt has a description of "the chemical activates expression of the protein."

Natural language inference tasks:

Listing 6: Prompt Format for generating the first sentence in NLI tasks with CLINGEN.

```
Suppose you need to create a set of [content]. Your task is to:
1. generate one sentence for a [content].
2. the [content] should be relevant to [topic],
3. The [content] should mimic the style of [style].
```

`[content]`: "health question" for MEDIQA-RQE, "claim" for MEDIQA-NLI, MedNLI and MQP, and "health news" for PUBHEALTH and HealthVer.

Listing 7: Prompt Format for generating the second sentence in NLI tasks with CLINGEN.

```
Suppose you need to create a pair of sentences for the [domain]
task with the label '[class_name]'. Given the [content]: '
[first_sentence]', Your task is to:
```

```
1. generate one short [content] about [topic] so that
[label_desc].
2. The [content] should mimic the style of the first sentence.
```

[domain]: "Question Entailment" for MEDIQA-RQE, "Natural Language Entailment" for MEDIQA-NLI and MedNLI, "Fact Verification" for PUBHEALTH and HealthVer, and "Sentence Similarity Calculation" for MQP.

[content]: "health question" for MEDIQA-RQE, "hypothesis" for MEDIQA-NLI, MedNLI, "evidence" for PUBHEALTH and HealthVer, and "sentence" for MQP.

[class_name]: the label name for this generated sample.

[label_desc]: the description of the selected label.

[first_sentence]: the first sentence we generate

G.3 Prompts for ZeroGen, DemoGen, ProGen

We use the same set of prompts for ZeroGen, DemoGen and ProGen, while DemoGen and ProGen have additional demonstrations augmented to the prompts. DemoGen uses the few-shot examples in the training set as demonstrations, and ProGen leverages feedbacks from previous rounds to iteratively guide the generation.

Named entity recognition tasks:

Listing 8: Prompt Format for NER tasks with baselines.

```
Suppose you need to create a dataset for [domain] recognition.
Your task is to generate a sentence about [domain] and output a
list of named entity about [domain] only.
```

[domain]: "disease" for BC5CDR-Disease and NCBI-Disease; "chemical" for BC5CDR-Chemical and CHEMDNER.

Medication attributes tasks:

Listing 9: Prompt Format for medication attributes tasks with baselines.

```
Suppose you need to create a dataset for clinical attributes
recognition. Your task is to generate a sentence about clinical
attributes, The Clinical Attributes you need to extract include "
Medication", "Dosage", "Route", "Frequency", "Reason", "Duration".
 For each attribute class, please return a list of attributes
within the class that occurs in the Sentence.
```

Text classification tasks:

Listing 10: Prompt Format for text classification tasks with baselines.

```
 Suppose you are a writer for [domain]. Your task is to give a
 synthetic [domain] about [class_name].
```

[domain]: "COVID-19 Literature" for LitCovid and "Cancer Document" for HOC.

[class_name]: the label name for this generated sample.

Relation extraction tasks:

Listing 11: Prompt Format for relation extraction tasks with baselines.

```
Suppose you need to generate synthetic data for the biomedical
[domain] task. Your task is to give a sentence about [class_name]
relation between [entity0] and [entity1] so that [label_desc].
```

[domain]: "Disease Gene Relation" for GAD, "Chemical Disease Relation" for CDR, and "Chemical Protein Relation" for ChemProt.

[entity0] and [entity1]: "disease" and "gene" for GAD, "chemical" and "disease: for CDR, and "chemical" and "protein" for ChemProt.

[class_name]: the label name for this generated sample.

[label_desc]: the description of the selected label. For example, the label "upregulator" in ChemProt has a description of "the chemical activates expression of the protein."

Natural language inference tasks:

Listing 12: Prompt Format for generating the first sentence in NLI tasks with baselines.

```
Suppose you need to create a set of [content]. Your task is to
generate one sentence for a [content].
```

[content]: "health question" for MEDIQA-RQE, "claim" for MEDIQA-NLI, MedNLI and MQP, and "health news" for PUBHEALTH and HealthVer.

Listing 13: Prompt Format for generating the second sentence in NLI tasks with baselines.

```
Suppose you need to create a pair of sentences for the [domain]
task with the label '[class_name]'. Given the [content]: '
[first_sentence]', Your task is to generate one short [content] so
 that [label_desc].
```

[domain]: "Question Entailment" for MEDIQA-RQE, "Natural Language Entailment" for MEDIQA-NLI and MedNLI, "Fact Verification" for PUBHEALTH and HealthVer, and "Sentence Similarity Calculation" for MQP.

[content]: "health question" for MEDIQA-RQE, "hypothesis" for MEDIQA-NLI, MedNLI, "evidence" for PUBHEALTH and HealthVer, and "sentence" for MQP.

[class_name]: the label name for this generated sample.

[label_desc]: the description of the selected label.

[first_sentence]: the first sentence we generate

H Additional Experimental Results

In this section, we present additional experimental results on every dataset in Tables 6, 7, 8.

	LitCovid	HOC	CDR			GAD			ChemProt
	F1	F1	P	R	F1	P	R	F1	F1
PubMedBERT$_{\texttt{Base}}$									
Supervised-Full	71.70	82.32	67.81	76.60	71.96	82.55	85.10	83.81	76.24
Supervised-Few	24.08	13.13	41.62	52.96	46.61	57.71	46.54	51.53	33.54
DA-Word Sub	36.49	44.98	40.50	46.20	43.16	51.15	32.10	39.45	31.82
DA-Back Trans	39.7	54.78	—	—	—	—	—	—	—
DA-Mixup	40.82	49.35	41.4	44.8	43.03	55.44	48.30	51.62	35.45
DA-Transformer	39.86	42.18	44.6	61.7	51.77	59.4	46.5	52.16	38.73
ZeroGen	50.50	67.90	38.82	**91.82**	54.57	84.38	80.68	82.49	54.46
DemoGen	57.65	70.52	46.9	83.3	60.01	93.14	80.19	86.18	56.18
ProGen	58.06	72.25	51.35	71.58	59.80	90.52	**85.14**	87.75	54.15
CLINGEN w/ KG	58.01	76.28	56.98	67.38	61.75	93.33	83.68	**88.24**	57.04
CLINGEN w/ LLM	**59.22**	**76.42**	**60.60**	66.35	**63.34**	**94.61**	78.17	85.61	**61.22**
PubMedBERT$_{\texttt{Large}}$									
Supervised-Full	74.59	85.53	72.31	74.88	73.57	84.95	88.75	86.81	78.55
Supervised-Few	22.59	13.13	42.27	67.51	51.99	57.58	90.07	70.25	35.80
DA-Word Sub	37.20	50.78	47.70	43.50	45.50	63.40	42.00	50.53	37.01
DA-Back Trans	40.50	61.46	—	—	—	—	—	—	—
DA-Mixup	40.03	53.45	43.34	73.50	54.53	62.20	59.93	60.52	37.87
DA-Transformer	38.95	49.86	50.70	31.60	38.93	59.80	57.76	58.76	40.66
ZeroGen	52.86	70.16	42.95	**80.67**	56.06	92.26	76.73	83.78	55.71
DemoGen	56.29	73.65	50.86	74.30	60.39	**96.85**	76.83	85.69	59.88
ProGen	54.71	75.31	50.36	76.08	60.60	91.11	85.63	88.29	58.79
CLINGEN w/ KG	55.81	77.71	60.45	65.04	62.66	94.30	**89.08**	**91.62**	60.12
CLINGEN w/ LLM	**57.07**	**78.14**	**67.13**	62.98	**64.99**	95.08	86.14	90.39	**63.05**

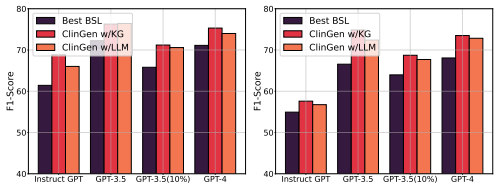
Table 6: Performance on single-sentence tasks evaluated by PubMedBERT$_{\texttt{Base}}$ and PubMedBERT$_{\texttt{Large}}$. **Bold** and underline indicate the best and second best results for each dataset, respectively (Same as below).

	MEDIQA-RQE	MEDIQA-NLI	MedNLI	PUBHEALTH		HealthVer		MQP
	ACC	ACC	ACC	ACC	F1	ACC	F1	ACC
PubMedBERT$_{\texttt{Base}}$								
Supervised-Full	77.15	79.01	81.43	65.16	62.96	70.00	68.02	75.70
Supervised-Few	57.51	40	36.40	28.30	23.70	30.55	30.49	55.70
DA-Word Sub	58.60	50.24	56.4	23.67	17.64	34.05	34.02	54.40
DA-Back Trans	59.16	49.92	53.82	30.70	23.32	33.60	32.76	55.80
DA-Mixup	57.71	49.38	53.47	31.45	24.45	34.11	33.78	58.20
ZeroGen	63.28	52.89	57.71	35.80	31.50	34.80	33.50	68.35
DemoGen	66.56	56.29	58.56	42.60	35.40	38.00	36.50	70.85
ProGen	65.94	57.28	59.49	38.70	33.10	36.72	35.97	69.30
CLINGEN w/ KG	**74.85**	58.03	61.80	44.60	36.80	43.05	42.06	72.17
CLINGEN w/ LLM	72.40	**64.44**	**64.89**	**48.50**	**40.60**	**44.50**	**42.32**	**73.31**
PubMedBERT$_{\texttt{Large}}$								
Supervised-Full	81.10	82.89	83.96	70.21	63.45	75.72	75.01	78.80
Supervised-Few	63.79	47.40	38.80	46.20	27.20	35.60	33.80	59.73
DA-Word Sub	64.26	51.20	57.53	35.60	31.60	35.41	32.29	55.30
DA-Back Trans	65.52	51.43	58.21	34.45	30.50	33.78	32.21	56.40
DA-Mixup	64.10	50.91	57.03	34.23	30.78	33.79	31.42	58.50
ZeroGen	67.26	60.74	62.42	42.50	33.30	39.74	38.90	72.69
DemoGen	69.22	62.97	64.55	44.50	36.80	40.72	40.57	74.37
ProGen	67.82	60.98	63.15	44.15	36.37	41.42	40.89	74.89
CLINGEN w/ KG	**79.92**	63.59	69.19	50.20	41.26	**47.03**	43.64	75.82
CLINGEN w/ LLM	77.36	**64.69**	**69.46**	**52.96**	**43.31**	46.05	**44.12**	**76.21**

Table 7: Performance on sentence-pair tasks evaluated by PubMedBERT$_{\texttt{Base}}$ and PubMedBERT$_{\texttt{Large}}$.

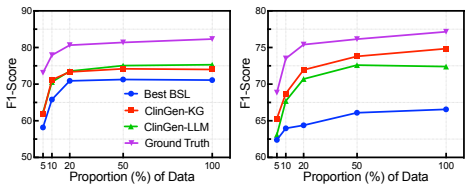
	BC5CDR-Disease			BC5CDR-Chemical			NCBI-Disease			CHEMDNER			CASI		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PubMedBERT$_{\texttt{Base}}$															
Full	83.84	87.92	85.83	92.22	91.74	91.98	87.54	89.92	88.71	91.84	92.45	92.14	—	—	—
Few	24.86	39.47	30.51	63.73	46.07	53.48	36.16	39.47	37.74	48.00	28.70	35.92	38.11	43.82	40.77
DA-Word Sub	35.34	39.54	37.32	63.13	52.52	57.34	53.40	36.70	43.50	47.45	33.15	39.03	40.25	47.65	43.64
DA-Mixup	36.13	42.90	39.23	66.43	50.54	57.41	56.57	26.48	36.07	52.40	27.53	36.10	42.37	48.96	45.43
DA-MELM	34.20	41.30	37.42	47.23	72.81	57.29	36.90	48.50	41.91	39.33	45.95	42.38	37.82	44.28	40.80
ZeroGen	55.60	39.10	45.91	73.20	82.85	77.73	56.25	45.98	50.60	**54.34**	52.93	53.63	52.80	49.53	51.11
DemoGen	63.10	48.44	54.81	76.40	81.65	78.94	57.65	49.08	53.02	54.00	53.77	53.88	58.15	56.84	57.49
ProGen	61.60	50.5	55.50	77.10	82.02	79.48	56.01	53.50	54.73	51.55	53.00	52.26	57.76	58.57	58.16
CLINGEN w/ KG	58.64	**63.02**	60.75	74.96	**85.45**	79.86	**62.62**	**56.62**	**59.47**	48.33	**69.28**	**56.94**	**71.75**	65.20	**68.32**
CLINGEN w/ LLM	**63.41**	58.83	**61.03**	**77.68**	84.33	**80.87**	62.58	50.59	55.95	51.40	58.77	54.84	68.19	**66.79**	67.48
PubMedBERT$_{\texttt{Large}}$															
Supervised-Full	86.77	85.92	86.34	92.80	92.94	92.87	87.97	90.09	89.02	92.23	92.48	92.35	—	—	—
Supervised-Few	25.52	45.85	32.79	61.40	54.41	57.69	44.86	40.12	42.35	43.40	34.60	38.50	41.30	45.02	43.08
DA-Word Sub	38.54	38.85	38.69	64.85	53.96	58.91	52.59	45.35	48.70	44.85	36.69	40.36	46.77	43.52	45.09
DA-Mixup	36.27	46.67	40.82	67.63	54.15	60.14	55.64	38.06	45.20	45.51	36.66	40.61	41.25	52.09	46.04
DA-MELM	33.40	41.61	37.06	53.80	66.71	59.56	44.20	57.40	49.94	36.40	47.41	41.18	43.36	45.78	44.54
ZeroGen	57.40	39.21	46.59	78.08	80.97	79.49	54.52	49.00	51.61	48.56	59.44	53.45	54.04	51.40	52.69
DemoGen	57.34	49.48	53.12	78.27	83.90	80.99	59.43	56.83	58.10	48.03	60.39	53.51	62.67	61.02	61.83
ProGen	60.34	54.13	57.07	**78.42**	82.94	80.62	60.02	55.28	57.55	50.40	59.64	54.63	57.21	63.70	60.28
CLINGEN w/ KG	54.28	**70.14**	61.21	77.88	86.32	81.88	**62.46**	**64.08**	**63.26**	47.03	**67.86**	**55.56**	70.96	**69.66**	**70.30**
CLINGEN w/ LLM	**61.05**	65.40	**63.15**	78.08	**86.98**	**82.29**	61.12	60.16	60.64	**50.92**	60.67	55.37	**71.61**	66.86	69.15

Table 8: Performance on token-classification tasks evaluated by PubMedBERT$_{\texttt{Base}}$ and PubMedBERT$_{\texttt{Large}}$.



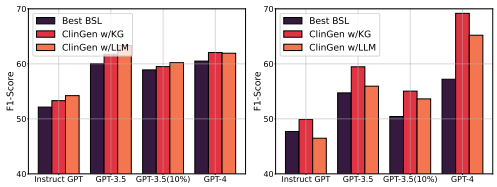
(a) HOC (b) MEDIQA-RQE

Figure 7: Different generators at `Base`.



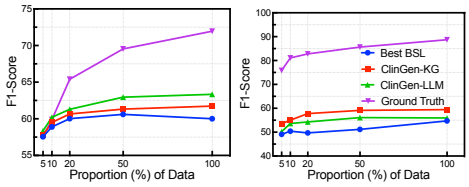
(a) HOC (b) MEDIQA-RQE

Figure 8: Different proportion of data at `Base`.



(a) CDR (b) NCBI-Disease

Figure 9: Different generators at `Base`.



(a) CDR (b) NCBI-Disease

Figure 10: Different proportion of data at `Base`.

I Ablation and Parameter Studies

Effect of Different LLM Generators. To investigate the impact of various LLMs on CLINGEN, we leverage other models in the GPT-family as the text generator. Specifically, we utilize InstructGPT (`text-curie-001`) [43] and GPT-4 [41]. Note that we only generate 500 samples in the GPT-4 setting due to budget constraints, but we provide the results of GPT-3.5 with same amount of synthetic samples for a fair comparison. From Figure 7 we observe that CLINGEN generally outperforms the best baseline in all settings. Additionally, we observe generally improved performance with larger models, as they often have better capabilities to follow our designed instructions for the given prompts. See Appendix I for more figures.

Effect of Size of Synthetic Data. In Figure 8 (and more in Appendix I), we study the effect of the size of synthetic data. The result shows that CLINGEN consistently outperforms the best baseline, using only around 10% of the synthetic examples. This illustrates that incorporating domain knowledge and increasing the diversity of the prompts could be an effective way to improve the sample efficiency, and narrow the gap between the performance of synthetic and ground-truth dataset.

Figure 9 and 10 show the effect of different generators and the effect of the proportion of data on two additional datasets, respectively. Overall, our method generally outperform the best baseline. One interesting finding for the NCBI-Disease dataset is that CLINGEN performs worse than the best on one variant. We hypothesize that it is because this task involves more complex input and output, potentially posing a challenge for moderate-size LLMs to follow the instructions.

Besides, as few-shot sample selection is important for the final performance, we show the performance of different 3 random seeds (with different seed examples/training process) in table 9, and observe that our method CLINGEN generally outperforms the baselines with non-negligible margins, which indicates the robustness of CLINGEN as it does not rely on a specific subset of few-shot training examples to perform well.

J Additional Quality Analysis

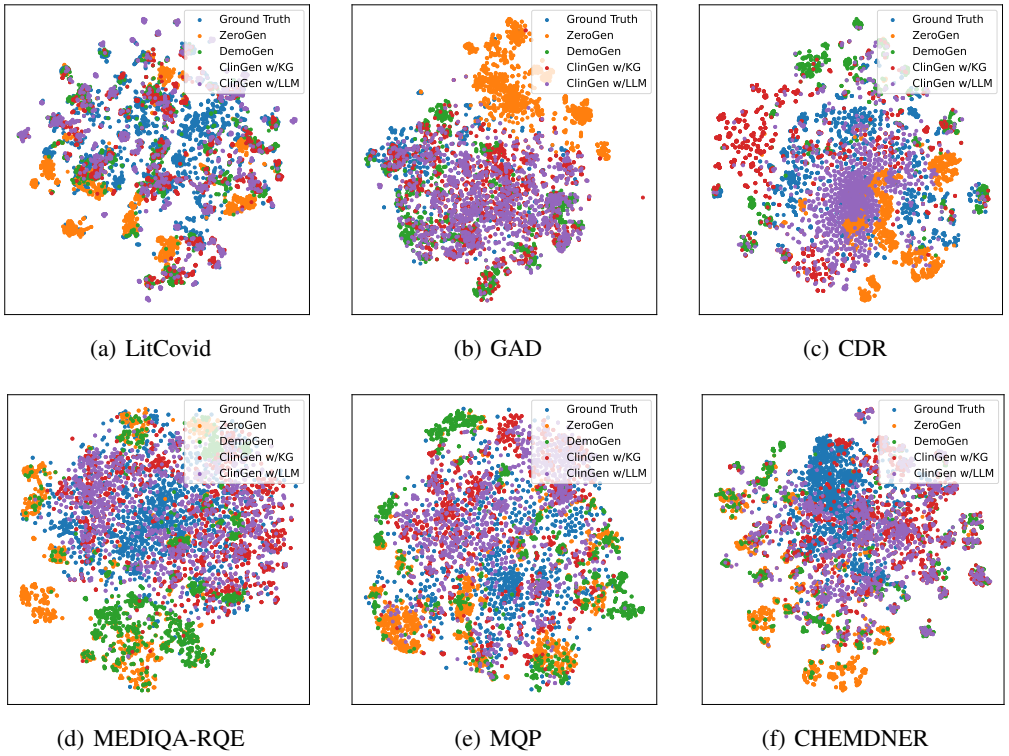
We present additional quality analysis of the synthetic dataset with t-SNE plots in Figure 11 and the regularized entity frequencies in Figure 12.

K Monetary Cost

We display the monetary cost of CLINGEN for calling the OpenAI APIs, with a comparison with prompting GPT-3.5 for direct inference and DemoGen. From the values shown in Figure 10, we

	HOC			CDR			MEDIQA-RQE			NCBI-Disease		
	Best Baseline	CLINGEN-KG	CLINGEN-LLM	Best Baseline	CLINGEN-KG	CLINGEN-LLM	Best Baseline	CLINGEN-KG	CLINGEN-LLM	Best Baseline	CLINGEN-KG	CLINGEN-LLM
1	70.04	74.30	77.30	61.52	61.66	63.34	68.30	76.85	74.50	56.12	60.22	54.51
2	75.30	79.73	73.63	60.69	63.77	64.66	64.20	71.80	71.19	54.19	60.64	57.81
3	71.41	74.81	78.33	57.82	59.79	62.02	67.18	75.90	71.51	53.85	57.52	55.50

Table 9: Performance with Different Random Seeds using PubMedBERT$_{\texttt{Base}}$.



(a) LitCovid (b) GAD (c) CDR

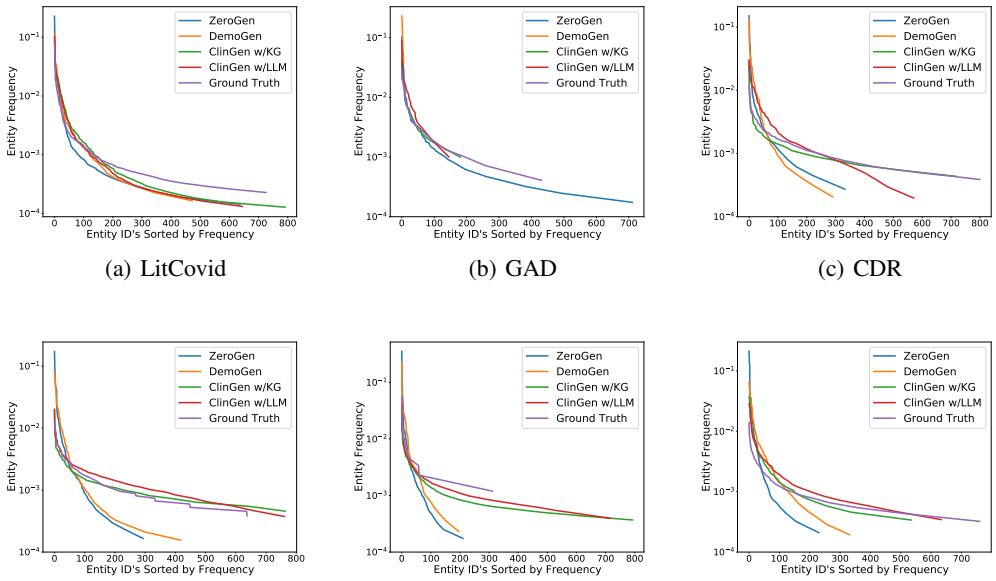
(d) MEDIQA-RQE (e) MQP (f) CHEMDNER

Figure 11: The t-SNE plots of datasets generated by CLINGEN, ZeroGen and DemoGen compared with the ground truth.

observe that inference via GPT-3.5 generally has a higher cost, as it needs to input all the testing samples for prompting. In contrast, DemoGen has a relatively lower cost, because it does not include the topics and writing styles to the prompts as CLINGEN does.

	HOC	GAD	ChemProt	MEDIQA-RQE	PUBHEALTH	NCBI-Disease	CASI
GPT-3.5 Inference	1.09	1.05	5.75	2.15	2.80	0.90	1.30
DemoGen	0.59	0.66	1.35	0.81	0.92	1.12	1.28
CLINGEN w/ KG	0.65	0.73	1.47	0.86	1.01	1.41	1.55
CLINGEN w/ LLM	0.72	0.84	1.51	0.90	1.34	1.49	1.62

Table 10: The average cost (in US dollars) of running CLINGEN on various datasets per 1000 samples, compared with prompting GPT-3.5 for inference and DemoGen.



(a) LitCovid (b) GAD (c) CDR

(d) MEDIQA-RQE (e) MQP (f) CHEMDNER

Figure 12: The regularized entity frequencies of datasets generated by CLINGEN, ZeroGen and DemoGen compared with the ground truth in log scale.