International Journal of Information Technology Vol. 29 No. 1 2023

Generative LLMs for Synthetic Data Generation: Methods, Challenges and the Future

Xu Guo[a] and Yiqiang Chen[b]

[a]Nanyang Technological University, Singapore
[b]Institute of Computing Technology, Chinese Academy of Sciences, China

Abstract

The recent surge in research focused on generating synthetic data from large language models (LLMs), especially for scenarios with limited data availability, marks a notable shift in Generative Artificial Intelligence (AI). Their ability to perform comparably to real-world data positions this approach as a compelling solution to low-resource challenges. This paper delves into advanced technologies that leverage these gigantic LLMs for the generation of task-specific training data. We outline methodologies, evaluation techniques, and practical applications, discuss the current limitations, and suggest potential pathways for future research.

Keywords: Generative AI, Synthetic Data Generation, Large Language Models..

I. Introduction

The introduction of Transformer [78] in 2017, followed by groundbreaking LLMs like OpenAI's GPT [6] and Google's BERT [16], marked the beginning of a new era in language understanding and generation. More recently, generative LLMs (e.g., GPT-3[37], LlaMa[77] and ChatGPT[59]) have propelled this evolution to unprecedented heights, seamlessly converging with Generative AI and heralding a fresh era in the realm of synthetic data generation[53, 52, 82, 21, 83, 86, 10].

The origins of Generative AI can be traced back to pivotal models such as Generative Adversarial Networks[24] (GANs) and Variational Autoencoders[36] (VAEs), which demonstrated the ability to generate realistic images and signals[80]. However, it wasn't until the advent of LLMs in recent years that Generative AI truly began to flourish. The convergence of Generative AI and LLMs in the realm of synthetic data creation represents not merely a technological advancement, but a profound paradigm shift in our approach to data creation and the training of AI models.

Why do we need synthetic data? The necessity for synthetic data arises from the inherent limitations of general-purpose Large Language Models (LLMs) in specialized and private domains, despite their significant achievements across various benchmarks. For instance, ClinicalBERT[33], adapted from BERT through pre-training on clinical texts, demonstrates superior performance in predicting hospital readmissions compared to the original BERT[15], which was trained on Wikipedia and BookCorpus[91] text data. This highlights a crucial challenge: specialized domains often rely on domain-specific data that is not readily available or open to the public, thereby underscoring the importance of synthetic data in bridging these gaps.

Synergy between LLMs and synthetic data generation. This synergy is pivotal in addressing data scarcity and privacy concerns, particularly in domains where real data is either limited or sensitive. By generating text that closely mirrors human language, LLMs facilitate the creation of robust, varied datasets necessary for training and refining AI models across various applications, from healthcare[65], eduction[57] to business management[70]. Moreover, this collaboration opens new avenues for ethical AI development, allowing researchers to bypass the biases and ethical dilemmas often inherent in real-world datasets.

Other related survey papers. Comprehensive surveys for Generative AI and LLMs exist, each revisits related works from a different perspective: Generative AI surveys provide a holistic view of this area starting from Generative Adversarial Networks (GANs) to ChatGPT [7] and models developed for synthetic data generation in the past decade [3], with a special focus on text-to-image [87] or text-to-speech [88] generation as well as practical applications in Education [1] and Healthcare [84]; Surveys for LLMs provide systematic categorization [66] for NLP tasks [56] and methods to adapt these LLMs to specific domains [25] through model optimization and personalization perspectives [28]. Surveys on LLMs for text generation [41] focus on developing generative LLMs including model architecture choices and training techniques and do not contain gigantic LLMs released in the past two years. Unlike these survey papers, this paper mainly focuses on recent technologies that employ generative LLMs *without training* them for synthetic training data generation and elicit their potential impact on practical adoption.

II. Generating synthetic training data from LLMs

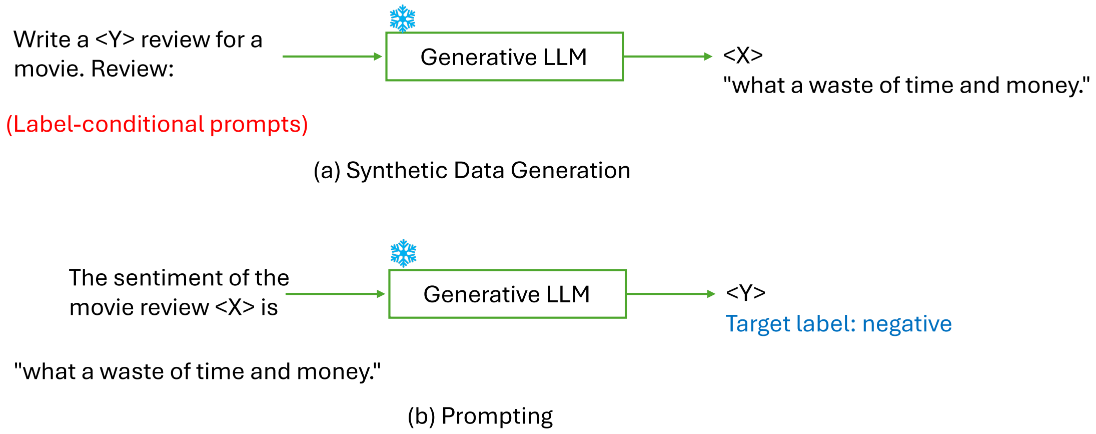
Write a <Y> review for a movie. Review:
(Label-conditional prompts)
Generative LLM
<X>
"what a waste of time and money."
(a) Synthetic Data Generation
The sentiment of the movie review <X> is
"what a waste of time and money."
Generative LLM
<Y>
Target label: negative
(b) Prompting

Fig 1 A general comparison between using LLMs for label-specific synthetic data generation (a) and label words prediction (b). In both cases, the LLMs are frozen and a task-related prompt is provided to condition the LLMs for task adaptation. $\langle X \rangle$ represents the text data and $\langle Y \rangle$ represents the label words.

Figure 1 shows the major difference between using generative LLMs for synthetic data generation and the predominant Prompting technique [6, 68] that directly applies LLMs for label prediction. In short, Prompting requires deploying the LLM model in practice to predict the label words $\langle Y \rangle$ (e.g., negative) from the input text data $\langle X \rangle$ with additional constraints from the prompt, e.g., “the sentiment of the movie review” indicates that the context is a movie review and the label shall describe its sentiment. On the contrary, synthetic data generation requires LLMs to generate text data $\langle X \rangle$ based on label-conditional prompts. It is the synthetic data distilled from LLMs rather than the LLMs themselves that will be applied in downstream applications, enabling more diverse and unlimited use cases based on synthetic data. Table 1 lists the newly emerging methods for generating task-specific training data from LLMs proposed in the past two years.

Method	Generator	Classifier	Benchmark
ZeroGen [82]	GPT2-XL [67]	LSTM[29] DistilBERT [73]	SST-2[74], IMDb[49], QNLI[69] RTE[14], SQuAD[69] AdversarialQA[2]
ZeroGen$^+$ [21]	GPT2-XL[67]	LSTM[29] DistilBERT [73]	IMDb[49], SST-2[74], Amazon[50] Rotten Tomatoes[63], Yelp[89] Subj[62], AGNews[89], DBpedia[89]
SuperGen [52]	CTRL[35]	COCO-LM[55] RoBERTa[46] GPT-2[67]	GLUE[79]
FewGen [53]	CTRL[35]	RoBERTa[46]	GLUE[79]
ReGen [86]	Condenser[22]	RoBERTa[46]	AGNews[89],DBpedia[89], MR[63] NYT[54], Yahoo[89], Amazon[50] Yelp[89], SST-2[74], IMDb[49]
ProGen [83]	GPT2-XL[67]	LSTM[29] DistilBERT [73]	SST-2[74], IMDb[49], Elec[50] Rotten Tomatoes[63], Yelp[89]
AttrPrompt [85]	ChatGPT[59]	BERT[16] DistilBERT [73]	NYT[54], Amazon[5] Reddit[23], StackExchange[23]
MixPrompt [10]	FLAN-T5 XXL [12]	GODEL [64]	NLU++[9],TOPv2[11] CrossNER [47]

Table 1 Data generation methods. Generator refers to LLMs that are used for synthetic data generation. Classifier refers to small-scale models that are trained on the synthetic data. These methods are limited to NLP models and tasks.

A. *Prompt engineering*

Designing an informative prompt is the key to effective data generation with LLMs. A simple and straightforward approach is to embed the label information in the prompt to refrain LLMs from generating label-agnostic data as described in Figure 1 (a). However, due to the limited number of words in labels and the limited task information in the prompt, the data generated by LLMs still can be unrelated to the task and lack diversity, limiting the size of the synthetic dataset that can be generated from the same LLM. As such, more advanced prompt engineering techniques are expected to circumvent the limitations of traditional ones.

Attribute-controlled prompt. A clear definition for a specific task can be obtained by specifying a set of attributes. Take News classification as an example, one piece of News article can differ from another by providing the details of $\mathrm{location}$, topic, $\mathrm{text\ genre}$ and so on. Inspired by this, MSP [10] employs a mixture of attributes in the prompt template to obtain desired synthetic data. In AttrPrompt [85], authors show that such attribute-specific prompts can be directly extracted from ChatGPT and then applied to query ChatGPT for generating attribute-specific data. By expanding the simple class-conditional prompt with more attribute constraints, we can gather more diverse synthetic data from LLMs while ensuring relevance to the given task.

Verbalizer. The verbalizer technique was originally proposed to enhance $\mathrm{Prompting}$ performance, where the target label words are expanded with their neighbouring words that hold the same semantic meanings [13, 32]. This strategy can be directly utilized to promote diverse data generation by expanding the class-conditional prompt into a set of semantically similar prompts. Besides, the verbalizer values can be extracted from LLMs themselves. For example, MetaPrompt [71] first obtains an expanded prompt from ChatGPT and further applies the enriched prompt to prompt LLMs for data generation.

B. Parameter-efficient task adaptation

Parameter-efficient approaches in the era of LLMs generally refer to the tuning methods that only tune a small set of an LLM's parameters (e.g., bias terms [4], embeddings or last layer) or an extra set of parameters that are inserted to LLMs (e.g., Adapters [30, 44], Prompt Tuning [40, 26], Prefix Tuning [42] and LoRA [31]). In the tuning process, the parameters of the LLM backbone are not updated and only the small set of trainable parameters are learned on task-specific datasets to achieve domain adaptation [17]. The advantage of parameter-efficient methods is that they grasp new task information while retaining powerful pre-trained knowledge. For example, FewGen [53] demonstrates that by tuning a few set of prefix vectors prepended to the CTRL model (1.6 Billion parameters) on few-shot datasets, the PrefixCTRL can generate more task-related training data. Similarly, MSP [10] trains a set of soft prompt embeddings on few-shot task-specific training data and then applies the trained soft prompts to condition the FLAN-T5 [12] (T5[68] further trained on instruction tuning datasets) for text generation.

C. Measuring data quality

In ZeroGen [82], authors measured the quality of the generated data from three quantitative perspectives: $\mathrm{diversity}$, $\mathrm{correctnes}$, and $\mathrm{naturalness}$. Results suggest that the quality of synthetic data is lower than real data in terms of the three perspectives. To obtain high-quality synthetic data, ProGen [83] proposes to incorporate a quality estimation module in the data generation pipeline, where the firstly generated synthetic data are evaluated by a task-specific model that was trained on oracle data in advance. Then, the most influential synthetic samples are selected as in-context examples to prompt GPT2-XL [67] to generate a new set of synthetic data. Similarly, authors in [18] employ a pre-trained classifier to filter out hard samples from synthetic datasets.

D. *Training with synthetic data*

To mitigate noise contained in the synthetic datasets, the implementation of regularization techniques is crucial for stabilizing training with noisy datasets. Innovations like ZeroGen$^+$ [21] suggest the use of a small weight network trained through bilevel optimization to autonomously determine sample weights. Additionally, FewGen [53] incorporates a self-supervised training approach using temporal ensembling [39]. This method has been shown to offer superior performance enhancements compared to label smoothing [58] when training downstream classifiers on synthetic data, highlighting its effectiveness in dealing with the unique challenges posed by synthetic datasets. Other techniques such as gradual annealing [19] also demonstrates to be effective in enhancing the learning performance on synthetic data.

III. Applications

A. *Low-resource and long-tail problems*

Low-resource problems generally suffer from the lack of sufficient data and in some cases particularly impacted by long-tail classes in practice [76]. Traditional research has predominantly leveraged transfer learning [27, 26] to enhance performance in low-resource settings. Yet, these methods hinge on the availability of relevant source-domain datasets, which may not always be accessible. A primary challenge in merging the research directions of synthetic data generation and low-resource learning tasks is navigating the distribution disparity between real and synthetic data, as well as optimizing the use of synthetic data in training scenarios. For instance, temporal ensembling employed in FewGen[53], gradual learning used in CAMEL [19], and the innovative data selection techniques proposed in [18], all contributed notable performance improvements.

B. Fast inference and lightweight deployment

Finetuning pre-trained language models on downstream tasks has been the predominant approach starting from the release of BERT [15]. However, the growing size of these language models, while enhancing performance, imposes practical burdens on organizations requiring swift inference and prompt responses. The shift towards synthetic data generation opens up a realm of possibilities for downstream applications. By generating a curated synthetic dataset, it becomes feasible to train smaller, less complex models, as demonstrated in [82, 21, 83]. This approach not only facilitates easier deployment but also ensures faster inference, addressing the critical need for efficiency in real-world applications.

C. Medical Scenarios

Data augmentation. Synthetic data generation can help some medical tasks that lack sufficient data to train a strong predictive model. For instance, studies in [61] demonstrated that augmenting real datasets with synthetic chest radiograph images generated by latent diffusion models[72] can enhance classification performance. In medical language processing, Tang et al. (2023) [75] demonstrated that tailored prompts provided to ChatGPT can yield task-specific synthetic data, significantly boosting the performance in tasks like biological named entity recognition and relation extraction. Additionally, GatorTronGPT, as explored in Peng et al. (2023) [65], which involved training GPT-3 from scratch on a dataset amalgamating 277-billion words from English and clinical texts, exhibited remarkable proficiency in generating synthetic clinical text.

Missing value imputation. Medical data can be sparse in that patients may take different or do not take some examinations, leading to imbalanced attributes. Missing value imputation (MVI) methods are helpful in enhancing the density of medical attribute values [45]. Traditional MVI approaches typically involve random sampling from specified value ranges, as noted in Luo et al. (2022) [48], essentially serving as a form of random data augmentation for certain attributes. With the advent of multi-modal LLMs, Ozbey et al. (2023) [60] demonstrate that in cross-modality translation tasks, missing images under specific attributes can be effectively imputed using synthetic images generated from diffusion models. Such synthetic data, compared to traditional random imputation methods, offer more diverse information, thereby helping to mitigate the issue of overfitting in attributes with limited data.

IV. Challenges with Synthetic Data and Future Directions

A. Overcoming Data Limitations

Correctness and Diversity. In Section II., we summarized existing approaches for monitoring the data quality and promoting data diversity in generation. They demonstrated effectiveness but do not entirely solved the problem. The challenge of ensuring the quality and accuracy of the generated data still remains profound. As an inherent nature, LLMs may inadvertently propagate inaccuracies or biases present in their pre-training data [43, 38], leading to outputs that may not always align with factual or unbiased information. Additionally, the intra-class and inter-class data diversity and domain representativeness are a concern, especially in specialized or niche domains.

Hallucination. Synthetic data generated by Large Language Models (LLMs) can sometimes be not only inaccurate but completely fictitious or disconnected from reality, a phenomenon often referred to as ”hallucination” [34, 90]. For instance, image generation based on specific captions can result in outputs with unrealistic features, such as a soldier depicted with three hands, as noted in the studies [19] for cross-modality generation. This hallucination issue is frequently linked to the quality of the training data, particularly if it contains inaccuracies that the LLM then overfits during the pre-training phase. Therefore, there's a pressing need to develop new, more effective strategies to detect and address hallucination [81] in the context of synthetic data generation, ensuring the reliability and authenticity of the output.

B. Data privacy and ethical concerns

While synthetic data offers a way to leverage the power of AI without compromising individual privacy[20], the ethical implications of using synthetic data, particularly in sensitive domains, raise questions about privacy and consent, as the boundaries between real and synthetic data blur. Research in [8] demonstrates that it is possible to extract specific information from the datasets used in training LLMs. Consequently, there exists a risk that synthetic data generation might inadvertently reveal elements of the underlying training data [51], some of which might be subject to licensing agreements. Moreover, uploading data to LLM APIs also remains a data privacy concern. For instance, employing LLMs in clinical text mining poses significant privacy risks related to uploading patient information to LLM APIs [75]. This challenge necessitates a careful balance between leveraging the benefits of AI and respecting the confidentiality and privacy of individuals, particularly in healthcare and other sensitive areas.

V. Conclusion

This paper reviews recent research on utilizing generative LLMs for synthetic data generation. With a focus on gigantic LLMs which are fixed for inference, we summarize recent generation methods for synthesizing training data. Additionally, we introduce some practical training techniques for training downstream models on the synthetic data presuming the data quality is inadequate. Then, we introduce some application scenarios extending from general low-resource issues to more specialized medical contexts. Finally, we conclude by spotlighting the significant ongoing challenges in the realm of synthetic data and proposing potential avenues for future research.

References

[1] David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023.

[2] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020.

[3] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.

[4] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] 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*, 33:1877–1901, 2020.

[7] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[8] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021.

[9] Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States, July 2022. Association for Computational Linguistics.

[10] Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. Mixture of soft prompts for controllable data generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14815–14833, Singapore, December 2023. Association for Computational Linguistics.

[11] Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. Low-resource domain adaptation for compositional task-oriented semantic parsing. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 5090–5100, Online, November 2020. Association for Computational Linguistics.

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[13] Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. Prototypical verbalizer for prompt-based few-shot tuning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer, 2005.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics.

[17] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models, 2022.

[18] Zilin Du, Haoxin Li, Xu Guo, and Boyang Li. Training on synthetic data beats real data in multimodal relation extraction, 2023.

[19] Zilin Du, Yunxin Li, Xu Guo, Yidan Sun, and Boyang Li. Training multimedia event extraction with generated images and captions. *arXiv preprint arXiv:2306.08966*, 2023.

[20] Meiling Fang, Marco Huber, and Naser Damer. Synthaspoof: Developing face presentation attack detection based on privacy-friendly synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1061–1070, 2023.

[21] Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, WEIZHONG ZHANG, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[23] Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. Tweac: Transformer with extendable qa agent classifiers, 2021.

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[25] Xu Guo. Data-efficient domain adaptation for pretrained language models, 2023.

[26] Xu Guo, Boyang Li, and Han Yu. Improving the sample efficiency of prompt tuning with domain adaptation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3523–3537, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[27] Xu Guo, Boyang Li, Han Yu, and Chunyan Miao. Latent-optimized adversarial neural transfer for sarcasm detection. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5394–5407, Online, June 2021. Association for Computational Linguistics.

[28] Xu Guo and Han Yu. On the domain adaptation and generalization of pretrained language models: A survey. *arXiv preprint arXiv:2211.03154*, 2022.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[30] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of*

Machine Learning Research, pages 2790–2799. PMLR, 09–15 Jun 2019.

[31] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[32] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification, 2022.

[33] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[34] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[35] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[37] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[38] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24, 2023.

[39] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[40] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient

prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[41] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*, 2022.

[42] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics.

[43] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[44] Meizhen Liu, Xu Guo, He Jiakai, Jianye Chen, Fengyu Zhou, and Siu Hui. InteMATs: Integrating granularity-specific multilingual adapters for cross-lingual transfer. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5035–5049, Singapore, December 2023. Association for Computational Linguistics.

[45] Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine*, page 102587, 2023.

[46] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[47] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460, May 2021.

[48] Fei Luo, Hangwei Qian, Di Wang, Xu Guo, Yan Sun, Eng Sing Lee, Hui Hwang Teong, Ray Tian Rui Lai, and Chunyan Miao. Missing value imputation for diabetes prediction. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[49] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[50] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

[51] R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670, 2023.

[52] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with lan-

guage models: Towards zero-shot language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[53] 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.

[54] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6826–6833, Jul. 2019.

[55] Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114, 2021.

[56] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[57] Steven Moore, Richard Tong, Anjali Singh, Zitao Liu, Xiangen Hu, Yu Lu, Joleen Liang, Chen Cao, Hassan Khosravi, Paul Denny, et al. Empowering education with llms-the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education*, pages 32–37. Springer, 2023.

[58] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help?, 2020.

[59] OpenAI. Introducing chatgpt, 2023.

[60] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

[61] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[62] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain, July 2004.

[63] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[64] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. Godel: Large-scale pre-training for goal-directed dialog, 2022.

[65] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *arXiv preprint arXiv:2305.13523*, 2023.

[66] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[67] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[68] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[69] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[70] Nitin Rane. Role and challenges of chatgpt and similar generative artificial intelligence in business management. *Available at SSRN 4603227*, 2023.

[71] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm, 2021.

[72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[73] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[74] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a

sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[75] 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.

[76] Anthony Meng Huat Tiong, Junnan Li, Guosheng Lin, Boyang Li, Caiming Xiong, and Steven C. H. Hoi. Improving tail-class representation with centroid contrastive learning, 2023.

[77] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[79] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[80] Yan Wu, Jeff Donahue, David Balduzzi, Karen Simonyan, and Timothy Lillicrap. Logan: Latent optimisation for generative adversarial networks, 2020.

[81] Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat.

Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564, 2023.

[82] 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 Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[83] Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. ProGen: Progressive zero-shot dataset generation via in-context feedback. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[84] Ping Yu, Hua Xu, Xia Hu, and Chao Deng. Leveraging generative ai and large language models: A comprehensive roadmap for healthcare integration. In *Healthcare*, volume 11, page 2776. MDPI, 2023.

[85] 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. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[86] Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. ReGen: Zero-shot text classification via training data generation with progressive dense retrieval. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805, Toronto, Canada, July 2023.

Association for Computational Linguistics.

[87] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[88] Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2, 2023.

[89] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

[90] 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.

[91] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.