# Using Synthetic Data for Domain Adaptation of Language Models
NLU Project Proposal

**Shubhankar Ranade**
New York University
shubhankar.r@nyu.edu

**Akash Gupta**
New York University
aksg@nyu.edu

**Abhinav Gupta**
New York University
gupta.abhinav@nyu.edu

**Victor Cui**
New York University
vyc8567@nyu.edu

## 1 Motivation and Related work

NLP has shifted from designing task-specific architectures to using large scale task-agnostic pretraining architectures. These Pretrained language models (PLMs) (Brown et al., 2020, Devlin et al., 2019, Liu et al., 2019) have demonstrated remarkable human level performance on multiple downstream NLU tasks when fine-tuned on a large amount of task-specific training data. However, procuring such a large amount of data is often expensive and time consuming, or even impossible. Moreover, domain shifts between training and testing dataset is detrimental to model performance.

As a workaround, Domain Adaptation (DA) techniques allow us to fine-tune models using easily accessible data that is not in the target domain of the task (e.g. synthetic data), and allow models to generalize better on out-of-domain data. In recent years, many DA techniques (Daumé III, Ramponi and Plank, 2020), with data selection strategies (Gururangan et al.) and representation learning (Ganin et al., 2015), have demonstrated good performance in fine-tuning PLMs without sufficient data in target domain.

In addition, recent studies have revealed the intriguing few-shot learning potential of PLMs (Brown et al., 2020, Gao et al., 2021, Scao and Rush, 2021, Schick and Schütze, 2021) and their ability to leverage task-specific information. With the motivation to tackle the lack of target data, we propose a domain-adaptation technique in which we are augmenting target task few-shot training data and then fine-tuning the domain-adapted pretrained model (DAPT, as in Gururangan et al.) on the augmented dataset.

## 2 Proposed Approach

To employ domain adaptation in a few shot setting, we propose a two step process as below:

- **Data generation:** implement a data generation technique similar to SuperGen proposed in (Meng et al., 2022, Hu et al., 2017), where training data is generated using a unidirectional PLM guided by label-descriptive prompts extracted from the limited amount of task-specific training data. As SuperGen is compatible with any generator PLM, we plan to start from moderately-sized models like GPT-2, then generalize to other LMs.

- **Task-specific fine tuning:** utilize the synthetically generated data to fine-tune a domain-adapted pretrained model as described in (Gururangan et al.) on various downstream tasks in a few-shot setting.

We plan to compare a base RoBERTa model against a DAPT model (as in (Gururangan et al.)) fine tuned using SuperGen (Meng et al., 2022), all in a few-shot setting. We can also compare the best performing DAPT + TAPT model trained on full task-specific dataset against DAPT + fine-tuning using SuperGen methods to evaluate the generation process itself.

### 2.1 Datasets and Tools required

To measure the benefit of using synthetic data to complement domain adaptive methods, we plan to use all 4 different domain datasets mentioned in (Gururangan et al.) for pre-training. For target tasks, we are specifically interested in the biomedical domain and plan to start with the CHEMPROT (Kringelum et al., 2016) and RCT datasets (Dernoncourt and Lee), which are both publicly available. We also consider other tasks such as MEDNLI (Romanov and Shivade, 2018) for task diversity. For experimentation, we will use PyTorch for prototyping and compute resources from NYU Greene/GCP cluster.

## References


Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

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, 2019.

Hal Daumé III. Frustratingly easy domain adaptation.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey, 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, et al. Don't stop pretraining: Adapt language models to domains and tasks.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 2015. doi: 10.48550/ARXIV.1505.07818. URL https://arxiv.org/abs/1505.07818.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723, 2021.

Teven Le Scao and Alexander M. Rush. How many data points is a prompt worth? In *NAACL*, 2021.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. 02 2022.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *ICML*, 2017.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. Chemprot-3.0: a global chemical biology diseases mapping. *Database : the journal of biological databases and curation*, 2016:bav123, Feb 2016. ISSN 1758-0463. doi: 10.1093/database/bav123. URL https://pubmed.ncbi.nlm.nih.gov/26876982. 26876982[pmid].

Franck Dernoncourt and Ji Young Lee. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts.

Alexey Romanov and Chaitanya P. Shivade. Lessons from natural language inference in the clinical domain. In *EMNLP*, 2018.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118, 2020. URL https://arxiv.org/abs/2009.07118.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40, 2007.

Mohammad Sadegh Rasooli and Joel R. Tetreault. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733, 2015. URL http://arxiv.org/abs/1503.06733. version 2.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005. ISSN 1532-4435.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.