Effects of Adaptive Pretraining in Specialized Domains for Named Entity Recognition

Jack Lynam
Christopher Newport University
1 Avenue of the Arts
Newport News, Virginia, USA 23606
jack.lynam@cnu.edu

Sam Henry
Randolph-Macon College
114 College Ave
Ashland, Virginia, USA 23005
samuelhenry@rmc.edu

Abstract

Due to unique concepts, syntactic structure, and vocabulary of specialized domains, it is common to train specialized Language models (LMs) for their target domain. For example, BioClinicalBERT is a specialized LM designed for clinical applications. These specialized LMs are typically created starting with a foundation model (such as BERT-base) which has been pretrained for the general English domain, and then adapted to the target domain via additional pretraining. Alternatively, LMs may be pretrained from scratch on data from the target domain. Both techniques are extremely computationally expensive and as such, these specialized LMs are often publicly released for other researchers. For some domains, such as the biomedical domain there are many, similar models available, and as a developer, this raises the question, which pretrained LM should I choose? Alternatively, in novel domains for which no specialized LMs exist, it raises different questions: Is it worth the cost to pretrain a LM from scratch? Should I adapt a general English model instead? Should I just use a general English model without adaptive pretraining? This is a particularly salient question when considering a limited budget. i.e. Should I pay for compute time or for annotators to create a larger dataset. In this paper we compare results of nine LMs across nine datasets spanning the clinical, scientific, and biomedical-related social media domains. From these comparisons we make several conclusions that can simplify the hyperparameter-tuning process and inform researchers and developers in novel domains. Broadly, these are that the effects of adaptive fine-tuning are small. If an adapted model exists in your domain, choose the one most closely related to your task. If no model exists, using a foundation model is likely sufficient.

1 Introduction

Named Entity Recognition (NER) is the task of identifying salient terms in text, and is a core NLP task performed within these larger systems, and as such, improving NER is a vital step to improving their effectiveness. Transformer (Vaswani et al., 2017)-based language models (LM) have become the state-of-the art for many NLP tasks, and despite recent advancements in Large Language Models (LLMs), smaller Language Models (LMs) such as BERT (Devlin et al., 2018) remain the state-of-the-art for low-level tasks such as NER. Although smaller than the larger LLMs, these models are still extremely computationally expensive to train from scratch and as such are typically created for maximum utility within the General English domain. However, specialized domains such as the biomedical and clinical domains have unique concepts, language, and grammatical structure (Nadkarni et al., 2011) meaning these General English models may not perform optimally without domain-specific adaptation.

Evidence suggests that BERT models perform better when pretrained on data closely related to their target task (Lee et al., 2020; Alsentzer et al., 2019; Gu et al., 2021; Peng et al., 2019), and previous work has compared several LMs across multiple datasets. For example, Peng et al propose the Biomedical Language Understanding Evaluation (BLUE) benchmark and compare the effects of pretraining on PubMed and PubMed + MIMIC using Large and Small BERT models. Gu et al (Gu et al., 2021) propose the Biomedical Language Understanding & Reasoning Benchmark (BLURB) benchmark and compare six biomedical language models and Base BERT over various tasks including five NER datasets. These works serve as important biomedical benchmarks to compare biomedical LMs. We distinguish ourselves from these works in focusing on understanding the effects of adaptive pretraining to answer practical questions such as, what LM should I choose or is it worth it to adaptively pretrain an LM for a novel domain? To answer this, we collected and compared results of

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 199–208

nine different LMs over nine popular NER datasets in the clinical, scientific, and biomedical-related social media domains.

We find that while domain-specific pretraining does increase performance, the increase is less dramatic than expected, not across-the-board, and often not statistically significant. This is even true for LMs trained from scratch for the target domain. Given the extremely high computational costs, these results have important implications for researchers working in novel domains without pre-existing domain-specific LMs, and can help guide the choice of a LM for researchers and developers in domains with pre-existing domain-specific LMs.

2 Background

2.1 Training Language Models

LMs consist of an Encoder-Decoder architecture. In this architecture, the encoder learns effective numeric representations of the text, and the decoder transforms those numeric representations into an output useful for the target task. LMs are trained in a sequence of steps: (1) *Pretraining* - train the encoder to create numeric representations of text. (2) *Optional adaptive pretraining* - adjust those numeric representations for the target domain. (3) *Fine-tuning* - train a decoder (and usually also adjust the weights of the encoder) to perform well on a specific task and dataset.

Pretraining is a self-supervised step in which the LM is trained on massive corpora containing billions of words. The precise pretraining routine varies depending on the specific LM (Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020; He et al., 2020), however most follow a routine similar to BERT (Devlin et al., 2018). BERT is jointly trained using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is performed by masking words in a sentence, and using the LM to predict the masked word. NSP is performed by providing two text samples to BERT, and training it to predict whether the second sentence follows the first sentence.

Pretraining is by far the most computationally expensive step of LM development, and creating a LM from scratch often carries prohibitive costs. As such, pretrained LMs are typically created for maximum utility and therefore trained for the general English domain. These pretrained LMs are often released as foundation models which are used by researchers directly or adapted to other datasets via adaptive pretraining.

In the *optional adaptive pretraining* step, the LM is further pretrained (typically using the same pretraining routine) on domain-specific corpora. For example, BERT-Base which is pretrained for the general English domain may adapted to the biomedical domain by performing additional pretraining using PubMed abstracts. While less computationally expensive than initial pretraining, this is still extremely computationally expensive (often prohibitively expensive) and the adapted LMs are commonly released for other developers in their domain.

Fine-tuning is the last step in training a LM. In this step, a task-specific decoder is attached to the pretrained weights of the encoder portion of the LM and the full model is trained to maximize performance on the provided dataset. This is often the only step performed by developers, who will select a pretrained LM using online repositories such as HuggingFace[1], then train it on a specific dataset. Common wisdom suggests you select a LM most closely related to your target domain.

2.2 Biomedical Data Sources

Publicly available databases have been created to enable research in biomedicine. Two of the most popular text databases are PubMed, which contains biomedical literature from online journals and books, and MIMIC-III, which contains standardized, de-identified clinical records.

The PubMed (Wheeler et al., 2007) and PubMedCentral (Roberts, 2001) resources provided by the National Center for Biotechnology Information (NCBI) are repositories of scientific articles related to the biomedical sciences. PubMed contains over 40 million article abstracts while PubMedCentral contains nearly 12 million full-text articles.

The Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016) is a vast, freely-available database containing about 2 million de-identified clinical records. MIMIC-III data is sourced from critical care reports at Beth Israel Deaconess Medical Center, a hospital in Boston, Massachusetts. These clinical records report on a variety of topics that include laboratory testing, billing and administrative work, patient demographics, discharge summaries, radiology reports, chart write-ups, and pharmacotherapy.

[1] https://huggingface.co/

Domain	Model	General English		Scientific		Clinical		Social Media	
		Wiki	Books Corpus	PubMed full-text	PubMed abstracts	MIMIC all-types	MIMIC discharge	Twitter	Reddit
General English	BERT	x	x						
Scientific	Bi-B	x	x	x	x				
	Blue-B	x	x		x				
	PMed-B				x				
Clinical	BiClin-B	x	x	x	x	x			
	BiDis-B	x	x	x	x		x		
	BlueB+M	x	x		x	x			
Social Media	BT	x	x					x	
	BiRed-B	x	x	x	x				x

Table 1: Language models and their pre-training data sources.

2.3 Pretrained Language Models for the Biomedical Domain

The biomedical domain is a broad, well-studied domain and as such has many pretrained LMs available for download. We divide the biomedical domain into three main sub-domains: (1) *clinical* which relates to the medical aspects of biomedicine, (2) *scientific* which relates to biology, chemistry, genetic, and other scientific aspects of biomedicine, and (3) *social media* which relates to consumer and patient views and testimony related to biomedicine. For these experiments, we selected nine different LMs which cover these sub-domains. All selected LMs except PubMedBERT are domain-specific adaptations of the general English BERT-Base model, and all but BERT Tweet are pretrained in the same way as BERT (e.g. MLM and NSP). Uniquely, PubMedBERT is trained from scratch for the scientific biomedical domain. We describe each evaluated LM below. To make the tables more compact, we assigned abbreviations for each of these LMs, which are listed in bold. Table 1 summarizes their difference in training data sources.

BERT: BERT-Base is the original, general English model proposed by Devlin et al (Devlin et al., 2018). It was pre-trained for the MLM and NSP tasks using the Books Corpus (Zhu et al., 2015) and English Wikipedia.

Bi-B: BioBERT (Lee et al., 2020) is initialized from BERT-Base and adaptively pre-trained on PubMed abstracts and PubMed full-text articles from PubMed Central. In the biomedical domain, Bi-B has significantly outperformed BERT on several tasks including NER and relationship extraction (Lee et al., 2020; Gu et al., 2021), and was therefore used as the initial weights for the next two systems.

BiClin-B: BioClinicalBERT(Alsentzer et al., 2019) is initialized from Bi-B and further pre-trained on all clinical notes in the MIMIC-III database. BiClin-B has been shown to be slightly more effective for clinical tasks than Bi-B (Alsentzer et al., 2019).

BiDis-B: BioDischargeSummaryBERT (Alsentzer et al., 2019) is initialized from BioBERT and further pre-trained on clinical notes which are discharge summaries from MIMIC-III. BiDis-B shows improvement over Bi-B for tasks with corpora that deal specifically with discharge summaries (Alsentzer et al., 2019).

Blue-B: BlueBERT (PubMed) (Peng et al., 2019) is initialized from BERT-Base and further pre-trained on PubMed abstracts. Blue-B pre-training data (PubMed) is similar to that of Bi-B, but excludes full-text articles. Blue-B has been shown to perform similarly to Bi-B (Peng et al., 2019).

BlueB+M: BlueBERT (PubMed+MIMIC) (Peng et al., 2019) is initialized from BERT-Base and further pre-trained on PubMed abstracts and additionally clinical notes from MIMIC-III. Therefore, this LM's pre-training data is similar to that of BiClin-B, but excludes full-text PubMed articles. BlueB+M has been shown to outperform Bi-B on clinical tasks (Peng et al., 2019).

PMed-B: PubMedBERT (Gu et al., 2021) is pre-trained from scratch on PubMed abstracts. Unlike the other LMs it is not adapted from BERT Base. Theoretically, this should be hugely beneficial because the vocabulary of a LM is fixed and is determined from the initial pretraining corpus. Whereas LMs adapted from general English may need to represent specialized terms using subword tokens, PMed-B can represent specialized terms

natively because they are present in its specialized vocabulary. PMed-B was shown to perform comparably or better than LMs like Bi-B and Blue-B for biomedical tasks (Gu et al., 2021).

BT: BERTweet (Nguyen et al., 2020) is initialized from BERT Base and further pre-trained on Twitter data. BT is not adapted to the biomedical domain, but is instead adapted to the unique, short-form, highly abbreviated nature of Tweets. We include BT to try to determine what is more important, social media structure or the biomedical context. Instead of following the original BERT pre-training procedure, BT's adaptive pre-training is modeled after the RoBERTa procedure (Liu et al., 2019). This pre-training procedure is distinct in two ways: masked tokens are changed each epoch during MLM and no NSP is performed. BT was shown to outperform similar LMs for NER in the general social media domain (Nguyen et al., 2020).

BiRed-B: BioRedditBERT (Basaldella et al., 2020) is initialized from Bi-B and further pre-trained on text from health-related Reddit threads. BiRed-B has shown good performance for biomedical text mining in the social media domain (Basaldella et al., 2020).

3 Datasets

We selected nine datasets which, like the selected LMs span the clinical, scientific, and social media subdomains of biomedicine. The datasets are summarized in Table 2 and brief descriptions are provided below.

i2b2 2010: The i2b2 2010 Concepts, Assertions, and Relations challenge (Uzuner et al., 2011) aimed to create suitable NLP systems for the clinical domain. Its associated dataset consists of 426 de-identified clinical notes from Partners Healthcare and Beth Israel Deaconess Medical Center. For the NER challenge, the clinical notes were annotated for medical problems, treatments, and tests. The challenge organizers provide a split of 170 training notes and 256 test notes.

n2c2 2018: The n2c2 2018 Track 2 challenge (Henry et al., 2020) focused on developing systems to recognize adverse drug events (ADEs). The dataset consists of 505 de-identified clinical notes gathered from the MIMIC-III clinical care database. For the NER challenge, the data was annotated for entities related to medications and ADEs which include: drug, strength, form, dosage, frequency, route, duration, reason, and ADE. The challenge organizers provide a split of 303 training notes and 202 testing notes.

BC5CDR: The BioCreative V Track 3 Chemical-Disease Relation (CDR) challenge (Wei et al., 2016) was created to develop systems that identify chemical-disease interactions. The CDR corpus is composed of 1500 PubMed abstracts annotated for diseases and chemicals. The task organizers provide a split of 500 training, 500 validation, and 500 test abstracts.

BC7DCPI: The BioCreative VII Track 1 drug-protein and chemical-protein interactions challenge (Miranda et al., 2021) focused on identifying biochemical entities for drug discovery. This dataset is a collection of 5000 PubMed abstracts annotated for genes and chemicals. The organizers provide a split of 3500 training abstracts, 750 validation abstracts, and 750 test abstracts. Background and large-scale background collections of automatically-annotated abstracts were also released for further analysis.

NCBI: The NCBI disease corpus (Dogan and Lu, 2012) is a collection of PubMed abstracts designed to test the efficacy of disease recognition systems. In total, 793 PubMed abstracts were annotated for disease mentions. While the corpus maintainers do not provide a training-test split for the data, one provided by (Crichton et al., 2017) has seen use in biomedical benchmarks (Gu et al., 2021).

NLMChem: The NLM-Chem corpus (Islamaj et al., 2021) was created for the BioCreative VII Track 2 task on identifying chemical information from full-text scientific publications. It contains 150 full-text PubMed articles annotated for chemical mentions. The corpus maintainers provide a split of 80 training documents, 20 validation documents, and 50 test documents.

BC7Med: The BioCreative VII Track 3 challenge (Weissenbacher et al., 2021) focused on identifying medication and dietary supplement mentions from social media data. The dataset contains tweets from 212 different users' timelines annotated for drug name mentions. This dataset is large, yet heavily imbalanced with only 0.2% percent of the tweets containing medication mentions. The challenge organizers provide a split of 89,004 tweets for training, 38,149 tweets for validation, and 54,482 tweets for the test set.

COMETA: The Corpus of Online Medical EntiTies (COMETA)(Basaldella et al., 2020) contains biomedical entities from Reddit. It was originally intended to evaluate entity linking systems and con-

Dataset	Domain	Entity Types	Data Source
BC5CDR	Scientific	Disease, Chemical	PubMed abstracts
BC7DCPI	Scientific	Gene, Chemical	PubMed abstracts
NCBI	Scientific	Disease	PubMed abstracts
NLMChem	Scientific	Chemical	Full-text PubMed articles
i2b2 2010	Clinical	Problem, Treatment, Test	De-identified clinical notes
n2c2 2018	Clinical	Drug, Strength, Form, Dosage, Frequency, Route, Duration, Reason, ADE	De-identified clinical notes
BC7Med	Social Media	Medication	Twitter data
COMETA	Social Media	Miscellaneous Biomedical Entities	Reddit threads
ADEMiner	Social Media	ADE	Twitter data

Table 2: Description of datasets and their sources used in evaluation.

tains 20,015 health-themed posts from Reddit annotated for biomedical entities found within the SNOMED CT ontology (Donnelly et al., 2006). These entities represent a wide variety of biomedical concepts including symptoms, diseases, chemicals, genes, procedures, and more. To re-purpose this dataset for NER, we identified the text spans of the annotated entities and used them as the gold standard. Unique to this dataset (due to its repurposing), all entities within this dataset received a label of "Biomedical Entity" or None.

ADEMiner: The DeepADEMiner corpus (Magge et al., 2021) is a combination of three different datasets (SMM4H 2017 (Sarker et al., 2018), SMM4H 2019 (Weissenbacher et al., 2019), and SMM4H 2020 (Gonzalez et al., 2020)) created for identifying and normalizing ADEs from Twitter data. There are a total of 33,246 tweets that are annotated for ADE mentions. No official training-testing split is provided.

4 Methodology

4.1 Data Processing

LM training requires a maximum of 512 token length samples, so each dataset was processed (chunked) into samples. For some datasets, a sample was clearly defined. This is the case for i2b2 2010, BC7Med, COMETA, and ADEMiner, where annotations are contained on one line and therefore a sample is identified as a single line of text. For other datasets, there is no clear distinction between samples because the text is annotated on a character-level basis and no pre-defined sample deliminator is provided. For these datasets, which include n2c2 2018, BC5CDR, BC7DCPI, NCBI, and NLMChem, we processed these datasets into samples using the ScispaCy sentence splitter (Neumann et al., 2019) with the `en_core_sci_sm` model [2].

[2] https://allenai.github.io/scispacy/

All datasets were tokenized using each LM's specific tokenizer (as specified by HuggingFace's `Transformers.Autotokenizer` library).

4.2 Language Model Architecture and Training

NER was posed as a multi-class token classification problem and performed by predicting a set of labels for each token token in that sample. The set of labels corresponds to the entity types of that dataset plus a *None* class. The `argmax` is taken over the set of labels to convert to a one-hot label encoding for each token. For example, in the BioCreative 5 CDR dataset, a vector of `<Disease, Chemical, None>` probabilities is predicted for each token, and converted to a vector containing all zeros except for a single one indicating the class.

We first performed a preliminary hyperparameter search to determine the learning rate, batch size, embedding dropout, and decoder architecture. Manual analysis of these results and those reported in the literature (Devlin et al., 2018; Alsentzer et al., 2019; Michalopoulos et al., 2020) support that fine-tuning with a learning rate of 1e-5, batch size of 16, and no embedding dropout produces LMs that perform well for biomedical NER. Due to the prohibitive costs of performing separate hyperparameter searches for all LMs and all datasets, we reused these hyperparameters for each LM across every dataset. For similar reasons, we also fixed the neural network architecture across all datasets. All BERT models have a distinct task-specific decoder, meaning they predict only classes found within a particular dataset. The decoder is a single densely-connected layer with softmax activation.

4.3 Results Generation

Reported results are the micro-averaged F1 scores averaged over 5-folds of cross-validation. For datasets that provided training, test, and/or vali-

dation splits, we combined all data into a single set which we used for cross-validation. Each fold consists of 70% for training, 10% for early stopping and 20% for validation (testing). The data-split was randomized but consistent across all compared LMs. To maximize the number of training samples per fold, we used the early stopping split to determine the optimal number of training epochs, then we combined the training and early-stopping data into a single training dataset and created the final model used to predict on the validation set. Preliminary analysis indicated this technique moderately improved results. Early stopping monitored micro-F1 with a patience of 5. The weights of each LM were re-initialized between validation folds.

4.4 Statistical Significance

t-tests are a commonly used test of statistical significance (Demšar, 2006). They require a set of evaluation results which are typically collected over each round of cross-validation or over multiple datasets. Paired tests provide more statistical power for detecting differences than unpaired tests and therefore care should be taken to ensure test sets are identical for all systems during cross-validation. Briefly, a t-test assumes that the two methods perform equivalently (this is the null hypothesis), and any difference can be explained by random variation. To test this, the t-test compares two distributions and calculates a test statistic which is used to calculate a p-value which indicates the probability of accepting the null hypothesis. In this case, the probability that the two systems are equivalent. Therefore, low p-values indicate a low probability that the systems are equivalent, meaning they are statistically significantly different. α, a cut-off threshold for p is used to draw conclusions. When $p < \alpha$ one can conclude that the two systems are significantly different. $\alpha = 0.05$ is commonly used and indicates a 95% probability that the systems are different. Since we are testing for a difference in the better (greater than) or worse (less than) direction, a two-tailed test is used.

Although t-tests are commonly used, they are not well-suited for comparing results over cross-validation or over multiple datasets. This is because the t-test assumes that the scores are independent and normally distributed. The scores across cross-validation are not independent because the training sets of each round overlap and the results may not be normally distributed. When comparing across datasets the scores are independent, but are not normally distributed. Several corrections to t-tests have been proposed to account for the non-independence of results (Nadeau and Bengio, 2003; Grandvalet and Bengio, 2006; Santafe et al., 2015). These tests update the estimate of variance to account for the non-independence of samples. An alternative to the t-test is the Wilcoxon signed-rank test (Wilcoxon, 1945). This is a non-parametric test which does not assume the data is independent nor normally distributed. Like the t-test, Wilcoxon produces both a test statistic and a p-value which may be used to determine statistical significance. Because of the few assumptions of non-parametric tests, Wilcoxon tests are arguably preferable to t-tests, even corrected t-tests (Demšar, 2006; Rainio et al., 2024) for comparing across rounds of cross-validation and across datasets, however because they are non-parametric they require more samples than parametric tests to make sound statistical inference. Importantly, Wilcoxon cannot produce a p-value less than 0.05 with only 5 samples, meaning, in our study it is impossible to conclude any two systems are different when comparing across 5 rounds of cross-validation. Therefore, we use Grandvalet and Bengio's T-test for Cross-Validation (Grandvalet and Bengio, 2006) as described by Santafe et al. (Santafe et al., 2015) when comparing two systems across rounds of cross-validation and the Wilcoxon signed-rank test when two systems comparing across datasets. For the T-Test for Cross-Validation we use a $corr = 0.35$ because it is a middle value balancing over-estimating and under-estimating statistical significance. To compare multiple systems we use the Friedman test, which is a non-parametric test designed for comparing multiple systems. While it is possible to use repeated t-tests or Wilcoxon tests to determine differences between many systems, repeated tests increase the risk of error (Rainio et al., 2024; Demšar, 2006) across the whole family of tests (e.g. when $\alpha = 0.05$ there is a 95% chance of being correct, but with 2 tests there is a $95\% * 95\% = 90.25\%$ chance of being correct for all tests.

5 Results

Table 3 shows the micro-F1 performance averaged over 5-fold cross validation for all LMs and datasets. Each row indicates the results for a dataset and each column indicates the results for a LM. Bold values indicate the highest score for that dataset. Rows are grouped by dataset type; first sci-

Dataset	BERT Model								
	English	Scientific			Clinical			Social Media	
	BERT	Bi-B	PMed-B	Blue-B	BiClin-B	BiDis-B	BlueB+M	BT	BiRed-B
BC5CDR	0.9540	0.9576	0.9518	**0.9610**	0.9545	0.9551	0.9555	0.2049	0.9579
BC7DCPI	0.8255	0.8274	**0.8300**	0.8293	0.8256	0.8260	0.8254	0.1349	0.8233
NCBI	0.8790	0.8693	**0.8835**	0.8765	0.8727	0.8670	0.8470	0.2081	0.8727
NLMChem	0.9354	0.8527	**0.9691**	0.9392	0.8513	0.8517	0.9347	0.1395	0.8521
i2b2 2010	0.8848	0.8883	0.8861	0.8927	0.8895	0.8893	**0.8988**	0.1607	0.8885
n2c2 2018	0.9142	0.9176	**0.9224**	0.9141	0.9178	0.9196	0.9183	0.0290	0.9180
BC7Med	0.5844	0.5894	0.5865	0.6151	0.6096	0.6078	**0.6732**	0.2957	0.6151
COMETA	0.6111	0.6108	0.6092	**0.6127**	0.6077	0.6074	0.6065	0.1359	0.6122
ADEMiner	0.5353	0.5289	0.5085	0.5193	0.5113	0.5192	0.5269	0.3104	**0.5509**
Mean	0.7915	0.7824	0.7914	0.7955	0.7822	0.7826	**0.7985**	0.1799	0.7878

Table 3: Validation set micro-averaged F1 scores averaged over 5-fold cross-validation.

entific, then clinical, then social media. Columns are similarly grouped. The final row shows the average performance of each LM over all datasets. An important note is that these results are not necessarily directly comparable to other reported results, since these are the average micro-f1 of 5-fold cross validation, not results on the provided test set (if provided). From this table we can make several observations:

Observation 1: BT performs poorly on all datasets, even BC7Med and ADEMiner which consist of Tweets. Recall that BT was trained on general English Tweets and not on biomedical data. This implies that modeling biomedical language is more important than modeling the short-form structure of Twitter. Therefore, when choosing a LM it may be best to consider the domain first and the target task's structure second.

Observation 2: PMed-B performs the best for most datasets, but it doesn't have the highest average among all datasets. This is because it doesn't do as well as other LMs on social media datasets. We suspect this is related to its difference in vocabulary, but it may also be because it has never been trained for the general English domain which social media posts may most closely reflect. Furthermore, although PMed-B performs the best for most scientific datasets it is never statistically significantly better than Bi-B or Blue-B (the other scientific models), but it is statistically significantly worse than both Blue-B ($p = 0.0032$ and Bi-B ($p = 0.0099$) on BC5CDR.

Observation 3: The results for most datasets are similar no matter what LM you choose (except BT). For example, the scores for BC5CDR (excluding BT) range from 0.9518 to 0.9610, a difference of only 0.0092 = 0.92%. This pattern of similar performance is observed in nearly all datasets. For 4 out of the 9 datasets the difference between best and worst performing (excluding BT) is less than 1%, for 7 out of the 9, this difference is less than 5%. The only large differences are for NLMChem (11.78%) and BC7Med (8.88%). Furthermore, when comparing the average performance across all datasets (the Mean row), the difference between the best and worst performing (excluding BT) is only 1.63%. This observation is supported by statistical significance testing. Using Friedman's test to compare the performance of all LMs across all datasets reveals a statistically significant difference between the LMs ($p = 0.0017$), but when removing BT from the comparison, Friedman's test calculates a $p = 0.4061$ revealing no statistically significant difference between any system except BT when compared across all datasets.

Observation 4: BERT is the base model from which all LMs (except PMed-B) are adaptively pretrained. However, despite this, the change in performance is minimal (excluding BT) for most datasets; less than 2% for 7 out of 9 datasets, 3.37% for NLMChem, and 8.88% for BC7MED. However, when comparing one dataset at a time across cross-validation scores, we found that at least one (but never all) adaptively pretrained system significantly improved upon BERT. This was true for all datasets except for NCBI and ADEMiner. This indicates that although significant improvement is possible it is not guaranteed.

Observation 5: Common wisdom is to choose a LM most closely related to your target domain. However, the results in Table 3 show that P-MedB (a scientific LM) performs the best for n2c2 2018 (a clinical dataset). Similarly for the social media datasets of BC7Med and COMETA, the best performing systems are BlueB+M (a clinical LM) and PMed-B (a scientific LM) respectively. At first

glance this would imply that common wisdom is not always advisable, however comparing across folds of cross-validation for n2c2 2018 the p-value between P-MedB and BiDis-B (the best performing clinical model) was $p = 0.0686$, indicating no statistical significance between these systems. Similar results of $p = 0.5078$ and $p = 0.6720$ were found for BC7Med and COMETA respectively when comparing against BiRed-B (the best performing Social Media LM). Therefore, common wisdom stands. You should choose a LM that is most closely related to your domain.

Observation 6: Similar LMs perform similarly. Friedman's test from observation 3 revealed that no two LMs other than BT are statistically significantly different when compared across all datasets. Not surprisingly, using Friedman'a test for within category performance across within category datasets (e.g. clinical LMs - BiClin-B, BiDis-B, BlueB+M - across all clinical datasets - i2b2 2010, n2c2 2108) also showed no statistical significance. $p = 0.1690$ for biomedical systems on biomedical datasets, $p = 0.8948$ for clinical systems on clinical datasets. Using Wilcoxon within the social media datasets reveals a statistical significance $p = 0.0039$ due to BT's poor performance.

We also compared within-category systems across folds of cross-validation for each dataset using t-test for cross-validation and found little significant differences. There was a single significant difference between scientific models on scientific datasets which was on BC5CDR for which (as previously noted) PMed-B performed significantly worse than both Bi-B and Blue-B. Among clinical systems there were no statistically significant differences on n2c2 2018, but on i2b2 2010 we found that BlueB+M is statistically significantly better than BiClinB ($p = 0.0192$). Lastly, as expected for all social media datasets BiRedB is statistically significantly better than BT on all datasets (BC7Med $p = 0.0035$; COMETA $p = 0.0073$, ADEMiner $p =< 0.0001$).

6 Conclusions and Limitations

In this paper we performed a comparison of nine different LMs on nine biomedical datasets spanning the scientific, clinical, and social media subdomains on the task of NER. We confirmed that you should choose a LM most closely related to your target domain, that models trained on similar data will perform similarly, and that only moderate improvements are gained by adaptive pretraining. However, we were surprised by how little the differences between systems typically were, particularly with respect to BERT versus its adaptively pretrained counterparts, and most particularly by PMed-B's only slight improvement despite huge computational costs.

Based on our observations we can make a few suggestions:

1) As an NER developer in a specialized domain, follow common wisdom and select a model most closely related to your target task. It is unnecessary to generate and compare results across many different LMs. Their performance difference will very likely be statistically insignificant. This is a relief particularly because NER is often a basic task used within larger systems. It frees up time and effort to focus on other system components and other hyperparameter tuning.

2) As researcher in a novel domain for which no adaptively pretrained models exist, you can prototype using BERT. Performance gains by adaptively pretraining may improve results, but don't expect a dramatic change. If results are promising, then it may be worthwhile to adaptively pretrain a model, but if results are not promising, focus on developing a new technique or annotating more data.

3) Pretraining a model from scratch is not worth the costs for all but the most important or widely-used domains. It can improve performance but adaptively pretraining is a comparable method that is computationally more efficient.

We wonder if these results generalize to other lower-level tasks, such as relationship extraction or span detection, and if these results generalize to large language models (LLMs) such as LLaMa (Touvron et al., 2023). Since LLMs are recognized as being able to better generalize to new data, we feel this is likely and believe that these results suggest that using non-adapted foundation models may be sufficient. This is particularly salient since adapting pretrained LLMs is typically prohibitively expensive for most researchers and developers.

Lastly, we acknowledge limitations: (1) Significance tests were performed across a small number of samples which reduces their power of inference. Furthermore, since they compare across a distribution of F1 scores they don't take into account the number of samples in each dataset. (2) LM and dataset specific hyperparameter tuning could reveal more stark differences between LMs.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. *arXiv preprint arXiv:2010.03295*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):1–14.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Rezarta Islamaj Dogan and Zhiyong Lu. 2012. An improved corpus of disease mentions in PubMed citations. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 91–99.

Kevin Donnelly and 1 others. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279.

Graciela Gonzalez, Ari Z Klein, Ivan Flores, Davy Weissenbacher, Arjun Magge, Karen O'Connor, Abeed Sarker, Anne-Lyse Minard, Elena Tutubalina, Zulfat Miftahutdinov, and 1 others. 2020. Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*.

Yves Grandvalet and Yoshua Bengio. 2006. Hypothesis testing for cross-validation. *Montreal Universite de Montreal, Operationnelle DdIeR*, 1285.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, and 1 others. 2021. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, 8(1):1–12.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2020. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv preprint arXiv:2010.10391*.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Claude Nadeau and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning*, 52.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific reports*, 14(1):6086.

Richard J Roberts. 2001. PubMed Central: The GenBank of the published literature.

Guzman Santafe, Iñaki Inza, and Jose A Lozano. 2015. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, and 1 others. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Davy Weissenbacher, Siddharth Rawal, Arjun Magge, and Graciela Gonzalez-Hernandez. 2021. Addressing extreme imbalance for detecting medications mentioned in twitter user timelines. In *International Conference on Artificial Intelligence in Medicine*, pages 93–102. Springer.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 21–30.

David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, and 1 others. 2007. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl_1):D5–D12.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. 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.