SweClinEval: A Benchmark for Swedish Clinical Natural Language Processing

Thomas Vakili, Martin Hansson, and Aron Henriksson
Department of Computer and Systems Sciences
Stockholm University, Kista, Sweden
{thomas.vakili, martin.hansson, aronhen}@dsv.su.se

Abstract

The lack of benchmarks in certain domains and for certain languages makes it difficult to track progress regarding the state-of-the-art of NLP in those areas, potentially impeding progress in important, specialized domains. Here, we introduce the first Swedish benchmark for clinical NLP: *SweClinEval*. The first iteration of the benchmark consists of six clinical NLP tasks, encompassing both document-level classification and named entity recognition tasks, with real clinical data. We evaluate nine different encoder models, both Swedish and multilingual. The results show that domain-adapted models outperform generic models on sequence-level classification tasks, while certain larger generic models outperform the clinical models on named entity recognition tasks. We describe how the benchmark can be managed despite limited possibilities to share sensitive clinical data, and discuss plans for extending the benchmark in future iterations.

1 Introduction

The field of natural language processing (NLP) has seen several important breakthroughs in the past decade. Currently, the field is dominated by pre-trained transformers models (Vaswani et al., 2017) that can be used to solve a wide and – ideally – diverse set of tasks. The capabilities of these models have to a large degree been tracked through the use of *benchmarks*, significantly helping to drive progress in the area. These evaluation suites test how the models perform on different pre-defined tasks and allow for comparisons between models and approaches.

While there are many benchmarks available, there are also many potential uses for NLP that they do not cover. Frequently, evaluations rely on English data (Joshi et al., 2020; Søgaard, 2022). However, a model performing well on an English benchmark in no way guarantees similar performance if the language changes. Additionally, benchmarks such as GLUE (Wang et al., 2018) tend to focus on tasks formulated for general-domain data. With increasing calls for NLP to be applied to specific domains, such as the clinical domain, there is a pressing need for benchmarks that address these areas.

The clinical domain, in particular, suffers from a lack of datasets for evaluating NLP systems. One critical reason for this is the inherently sensitive nature of clinical data. There are multiple studies (Carlini et al., 2021; Nasr et al., 2023) demonstrating the potential risks of using sensitive data for machine learning – let alone sharing data in their raw form. That said, there are some widely used resources for clinical NLP. Prominent examples include the various versions of MIMIC (Johnson et al., 2022) and the i2b2 datasets (Murphy et al., 2010). Crucially, these datasets predominantly evaluate NLP systems on data in English or other higher-resourced languages.

In this paper, we introduce the first Swedish benchmark based on real clinical NLP data: *SweClinEval*. This benchmark consists of datasets built from electronic health records from the Health Bank (Dalianis et al., 2015) and includes a wide range of clinical tasks. These tasks include three different document-level sequence classification tasks and three token-level named entity recognition (NER) tasks. This introduction of SweClinEval includes nine different models, and future additions will be added to the benchmarks online leaderboard[1].

The evaluations presented in this paper show that many models targeting Swedish data per-

[1]The leaderboard of SweClinEval is available at: https://sweclineval.dsv.su.se

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 767–775
March 3-4, 2025

form strongly on our benchmark. However, the performances vary, and several interesting trends emerge from our results. These results highlight the importance of continuing to focus on domain-specific evaluations for languages other than English. Our results demonstrate the current state of Swedish clinical NLP, and the benchmark serves as an important tool for monitoring progress in this important NLP domain.

2 Related Research

The NLP community has seen impressive advances in the past few years with the advent of LLMs. Several new model architectures have been proposed since Vaswani et al. (2017) described the transformer, and new models are released at a rapid pace. These LLMs aim to be general-purpose models, with task-specific applications requiring only smaller adjustments in the form of fine-tuning or prompt engineering. In response to this new paradigm, there has been an increasing focus on creating benchmarks that capture the nuanced difference in performance in the growing plethora of models.

2.1 General-Domain Benchmarks

Benchmarks come with different objectives and designs. A prominent example is the GLUE (Wang et al., 2018) family of benchmarks. The original *General Language Understanding Evaluation* (GLUE) benchmark aimed to, as the name suggests, capture a wide range of capabilities that act as proxies for natural language understanding. As models have become more powerful, the NLP community has responded with more varied and difficult benchmarks. These include the SuperGLUE (Wang et al., 2019) benchmark that introduces more difficult tasks, and the XGLUE benchmark (Liang et al., 2020) that also examines the multilingual capabilities of models.

2.2 Swedish Benchmarks

The vast majority of papers at NLP conferences focus on English data (Søgaard, 2022), to the detriment of smaller and less well-resourced languages. The introduction of multilingual benchmarks such as XGLUE is in part a response to this dominance of English-only datasets.

Another development is the creation of language-specific benchmarks. For Swedish, this trend has materialized in the form of benchmarks such as the Superlim[2] (Berdicevskis et al., 2023) and OverLim[3] benchmarks. These benchmarks mirror the structure of the GLUE family of benchmarks, but use datasets that specifically use Swedish data.

An important benchmark, especially for the purposes of this paper, is the ScandEval (Nielsen, 2023) benchmark. This benchmark is multilingual but focuses mainly on the Scandinavian language family. LLMs for these languages have been found to benefit from training on shared datasets. The ScandEval benchmark was also used to determine which models to benchmark, as detailed in Section 3.2.

2.3 Clinical Benchmarks

The most commonly used benchmarks aim to measure general-purpose capabilities in a general-domain setting. However, many important applications of NLP are domain-specific. In this paper, we focus on NLP for clinical data, which has several domain-specific features. Due to the setting in which they are produced, clinical data are often riddled with domain-specific acronyms and terminology that can be harder for general-domain models to process (Dalianis, 2018). Furthermore, clinical datasets are difficult to share due to the inherently sensitive nature of the data.

Nevertheless, there have been efforts to create benchmarks that measure the clinical or biomedical capabilities of LLMs. BLURB (Gu et al., 2021) is a benchmark in the vein of GLUE and includes a wide range of clinical tasks. This benchmark highlighted the shortcomings of general-domain models and the benefits of using LLMs specific to the clinical domain. In contrast, the later Dr. Bench (Gao et al., 2023) benchmark shows that general-domain models can indeed outcompete domain-specific models on certain tasks. These diverging conclusions exemplify the need for diverse domain-specific benchmarks to monitor the progress of LLMs in the clinical domain.

A recent benchmark highly relevant for Swedish biomedical NLP is the *Swedish Medical Benchmark* introduced by Moëll and Farestam (2024). This benchmark is comprised of a selection of four datasets with multiple-choice questions. These datasets were collected from public

[2]Superlim is Swedish for super glue, a reference to the SuperGLUE benchmark.

[3]https://huggingface.co/datasets/KBLab/overlim

sources and probe LLMs for biomedical knowledge. A benefit of using publicly available data is that the data can be shared. On the other hand, such data are not representative of the types of clinical data and tasks encountered when creating, for example, a system interfacing with patient records.

The main contribution of this paper is the introduction of the SweClinEval benchmark. This benchmark is not only focused on the clinical domain, but is the first benchmark that monitors the state of Swedish clinical NLP using real electronic patient records for realistic clinical tasks.

3 Methods and Materials

Creating this first rendition of SweClinEval involved collecting resources for evaluation and deciding how to conduct the evaluations. This section describes the datasets used for the benchmark and the models that were tested, and how they were chosen. The design of the evaluations and the metrics used for comparing models are also described.

3.1 Datasets

The benchmark consists of six datasets that are part of the Health Bank (Dalianis et al., 2015) infrastructure[4] . The Health Bank consists of over 2 million Swedish electronic health records written between 2006 and 2014 from a range of different clinical units in Sweden. The datasets have been collected for more than a decade, either through manual annotation or by mining information from the Health Bank data. Three of the datasets are document-level classification tasks, and the other three are token-level NER tasks.

ICD-10 The Stockholm EPR Gastro ICD-10 Corpus (Remmer et al., 2021) is a document-level classification task where discharge summaries related to gastrointestinal patients are assigned high-level diagnosis code blocks. These 10 different code blocks encode information about what type of diagnosis was assigned to the patient. The task is a multi-label classification task, meaning that each document can be associated with more than one code block.

ADE The Stockholm EPR ADE ICD-10 Corpus (Vakili et al., 2022) is another document-level classification task that determines whether or not a discharge summary describes a patient suffering from an adverse drug event. This is a binary classification problem.

Factuality The Stockholm EPR Diagnosis Factuality Corpus (Velupillai, 2011; Velupillai et al., 2011) is the third document-level classification task. This manually annotated corpus assigns a *factuality* level to the diagnoses of each clinical note. These different levels describe the confidence with which a diagnosis was decided. The six different classes are: *Certainly Negative, Probably Negative, Possibly Negative, Possibly Positive, Probably Positive,* and *Certainly Positive*.

Factuality NER This version of the Stockholm EPR Diagnosis Factuality Corpus is a token-level NER task. The task involves assigning the same six labels to tokens in each document that indicate a diagnosis. The task is to both detect these diagnoses and assign them a factuality level. This version also includes an *Other* tag for clinically relevant information that is not indicating factuality.

Clinical Entity NER The Stockholm EPR Clinical Entity Corpus (Skeppstedt et al., 2014) is a manually annotated NER corpus that describes a task in which the model needs to identify clinically relevant terms. These are divided into four classes: *Diagnosis, Findings, Body Parts*, and *Drugs*. The model needs to detect tokens associated with these classes and assign them the correct labels.

PHI NER The final corpus used in the benchmark is the Stockholm EPR PHI Corpus (Dalianis and Velupillai, 2010). This corpus consists of patient records and has been manually annotated for named entities describing personally identifiable protected health information (PHI). Each instance of PHI is assigned one of nine classes: *First Name, Last Name, Age, Phone Number, Partial Date, Full Date, Location, Health Care Unit,* and *Organization*.

Additional statistics about the six datasets are listed in Table 1. None of the datasets have been

[4]This research has been approved by the Swedish Ethical Review Authority under permission no. 2019-05679.

adapted for use with prompt-style autoregressive language models. This limitation is reflected in the model selection for this paper and adapting the datasets for broader use is left to future iterations of SweClinEval.

3.2 Models

Nine different models were included for the experiments in this paper and are listed in Table 2. Two of these – SweDeClin-BERT and SweClin-BERT – were specifically created for use in Swedish clinical NLP and have previously shown strong performance on the datasets in SweClinEval (Vakili et al., 2024). Additionally, seven general-domain models known to perform well for Swedish data were included. These seven models were selected based on their performance in the ScandEval (Nielsen, 2023) benchmark.

The majority of the models are based on the BERT/RoBERTa architecture (Devlin et al., 2019; Liu et al., 2019). The RemBERT (Chung et al., 2020) and Multilingual E5 Large (Wang et al., 2024) models are based on their own transformer architectures. These two models also exhibit the greatest language diversity in their training data. The training data for the *RoBERTa Large* and *BERT Large* models from AI Sweden are also multilingual. These were trained using *The Nordic Pile* corpus (Öhman et al., 2023) which consists mainly of Scandinavian and English data.

Crucially, all nine models are encoder models. This is a limitation imposed by the nature of the datasets, as described in the previous section. It is possible to restructure datasets so that they can be used autoregressively. However, such a conversion would be non-trivial and is left for future research.

3.3 Evaluation Procedure

All nine models were trained and evaluated using the six datasets. To ensure a fair estimate of each model's performance, the evaluations were done using 10-fold cross-validation. This allowed us to calculate the average performance alongside the standard deviation, enabling a more fair comparison. The comparisons were based on the F_1 scores of each cross-validation.

For each fold in the cross-validation, models were trained for a maximum of three epochs. Early stopping was enabled, and the best-performing checkpoint was used to predict the test set in each fold. The F_1 scores used for the comparisons were based on the average score from each fold and the standard deviation. For the NER tasks, these were the token-level micro F_1 scores. The *PHI NER* task uses the IOB scheme to mark where an entity begins and ends, and this distinction was included in the evaluation. The document-level sequence classification tasks instead rely on F_1 scores weighted for the support of each class in the test set.

4 Results

Nine models were evaluated using 10-fold cross-validation for six different datasets, resulting in 540 evaluations. The average F_1 scores and their deviations are listed in Table 3.

For the sequence-level classification tasks, the highest average F_1 scores are consistently obtained using the domain-adapted models. The same is not true for the token-level NER tasks. For these tasks, the highest F_1 scores were obtained by the general-domain *RoBERTa Large* model from AI Sweden. However, the domain-adapted *SweDeClin-BERT* model has the second-highest average F_1 scores for the *Factuality NER* and *Clinical Entity NER* tasks.

The different average F_1 scores vary substantially between the best- and worst-performing models. Nevertheless, the standard deviations are large. This means that many of the averages are within a standard deviation of a competing model. This necessarily limits the analysis into which models are *best*, since randomness has a strong influence on the variability in the F_1 scores.

In addition to the predictive performance, Table 4 also lists the processing time of each model when performing inference. Unsurprisingly, the smaller models are faster to run. These figures are based on the HuggingFace implementations of each model running on an *Nvidia RTX A5000* GPU. Although the exact inference time will depend on the hardware available, the number indicate the relative cost of running these model in a production environment.

5 Discussion

A few trends emerge from the results in the previous section. There are also some limitations and pointers to future work that are important to discuss. However, we begin by discussing the findings from our results.

As previously mentioned, the highest average F_1 scores in the sequence classification tasks are

Task	Type	Classes	Documents	Tokens
ICD-10	Classification	10	6,062	930,550
ADE	Classification	2	21,725	931,778
Factuality	Classification	6	3,710	102,223
Factuality NER	NER	7	3,822	286,205
Clinical Entity NER	NER	4	3,120	178,672
PHI NER	NER	9	29,560	282,820

Table 1: Six different datasets were used in the benchmark evaluation. Three of these are NER tasks and three are sequence classification tasks. This table lists the datasets alongside their size, the number and classes, and the types of classification they target.

Model	Parameters	Paper
SweDeClin-BERT	125 M	(Vakili et al., 2022)
SweClin-BERT	125 M	(Lamproudis et al., 2021)
KB-BERT Base	125 M	(Malmsten et al., 2020)
AI Nordics BERT Large	335 M	N/A[5]
AI Sweden RoBERTa Large	355 M	N/A[6]
AI Sweden BERT Large	369 M	N/A[7]
KB-BERT Large	370 M	N/A[8]
Multilingual E5 Large	560 M	(Wang et al., 2024)
RemBERT	576 M	(Chung et al., 2020)

Table 2: In this initial edition of the SweClinEval benchmark, nine different models were evaluated. All models are encoder models, and they are listed here in order of parameter count. When available, the paper that introduced the model is listed. SweDeClin-BERT and SweClin-BERT are the only models created specifically for Swedish clinical NLP.

achieved by the domain-adapted models. This indicates that, at least for these tasks, domain adaptation results in better performance on clinical NLP tasks. On the other hand, this finding is not as clear when examining the NER tasks. While the domain-adapted models perform competitively, the best-performing model on all three NER tasks is AI Sweden's *RoBERTa Large* model.

Crucially, the models differ greatly in size. The smaller models are around three times smaller than the medium-sized models, and more than four times smaller than the largest models. The comparatively strong performance of the domain-adapted models, which are both small, is more impressive when seen from this perspective. Domain adaptation seems to allow smaller models to compete with larger counterparts. Naturally, this leads to the question of whether this finding holds true for larger models, too. The two clinical models are initialized from *KB-BERT Base*, and an interesting direction for future work could be examining if initializing from larger models produces analogous results. The *RoBERTa Large* model from AI Sweden would be an interesting candidate, given its strong performance on the NER tasks. In any case, the benefits from domain adaptation align with many previous studies (Gu et al., 2021; Lamproudis et al., 2021).

Perhaps somewhat surprisingly, parameter count itself does not seem to be a determining factor in what models are the strongest. This is not only the case when comparing domain-adapted and general-domain models. For example, *KB-BERT Base* and *KB-BERT Large* were both trained by the same organization, and are from the same model family. The main difference between the

[5] https://huggingface.co/AI-Nordics/bert-large-swedish-cased

[6] https://huggingface.co/AI-Sweden-Models/roberta-large-1160k

[7] https://huggingface.co/AI-Sweden-Models/bert-large-nordic-pile-1M-steps

[8] https://huggingface.co/KBLab/megatron-bert-large-swedish-cased-165k

Model	Size	ICD-10	Factuality	ADE
		Classification	Classification	Classification
SweDeClin-BERT	S	0.832±0.011	**0.735±0.018**	**0.203±0.022**
SweClin-BERT	S	**0.836±0.014**	0.731±0.021	0.196±0.014
KB-BERT Base	S	0.801±0.015	0.671±0.017	0.185±0.012
AI Nordics BERT Large	M	0.811±0.012	0.657±0.025	0.192±0.013
AI Sweden RoBERTa Large	M	0.816±0.018	0.594±0.126	0.159±0.028
AI Sweden BERT Large	M	0.816±0.012	0.654±0.032	0.167±0.057
KB-BERT Large	M	0.801±0.013	0.683±0.019	0.190±0.011
Multilingual E5 Large	L	0.824±0.013	0.525±0.074	0.192±0.015
RemBERT	L	0.823±0.010	0.379±0.059	0.149±0.050

Model	Size	Factuality	Clinical Entity	PHI
		NER	NER	NER
SweDeClin-BERT	S	0.623±0.024	0.766±0.034	0.945±0.012
SweClin-BERT	S	0.610±0.018	0.754±0.038	0.938±0.014
KB-BERT Base	S	0.600±0.025	0.743±0.039	0.941±0.025
AI Nordics BERT Large	M	0.612±0.026	0.721±0.039	0.948±0.010
AI Sweden RoBERTa Large	M	**0.641±0.011**	**0.779±0.036**	**0.965±0.009**
AI Sweden BERT Large	M	0.513±0.185	0.738±0.038	0.854±0.285
KB-BERT Large	M	0.552±0.025	0.697±0.046	0.936±0.012
Multilingual E5 Large	L	0.603±0.019	0.511±0.339	0.608±0.037
RemBERT	L	0.417±0.026	0.600±0.075	0.947±0.011

Table 3: Nine encoder models were evaluated for sequence classification using six different clinical tasks. Three of the tasks were sequence classification tasks, and three were token-level NER tasks. The performance is summarized using F_1 with standard deviations. The highest F_1 of each task is bolded, and the second highest is underlined. Models are ordered according to ascending parameter count as listed in Table 2 and categorized as ***S**mall, **M**edium,* or ***L**arge* models.

Model	Sequence	NER
SweDeClin-BERT	2.86 ms	2.85 ms
SweClin-BERT	2.86 ms	2.84 ms
KB-BERT Base	2.88 ms	2.87 ms
AI Nordics BERT Large	5.60 ms	5.56 ms
AI Sweden RoBERTa Large	6.91 ms	6.05 ms
AI Sweden BERT Large	5.60 ms	5.56 ms
KB-BERT Large	8.76 ms	8.67 ms
Multilingual E5 Large	6.08 ms	6.03 ms
RemBERT	9.38 ms	9.36 ms

Table 4: The different models used in the benchmark use different architectures and are of different sizes. This table lists the time of each model for inference on one sample, both for sequence classification and NER.

models is that the larger model consists of more parameters and was trained using a much larger corpus. Nevertheless, *KB-BERT Base* actually outperforms its larger counterpart in some cases. While the large standard deviations call for cautious interpretations of the results, it is at least clear the larger model is not outperforming its smaller competitor.

On the other hand, parameter count clearly influences the inference speed of the models, as indicated in Table 4. While this is not surprising, it is worth mentioning. Other benchmarks, such as the GLUE benchmark, do not always present this information. However, inference speed can be important in practice, especially when differences in performance are small. Smaller and faster models require less expensive hardware, which can be important in cases where it is not possible to use cloud providers to run the models. This is frequently the case for clinical uses, due to the sensitivity of clinical data.

6 Conclusions

In this paper, we present SweClinEval – the first Swedish benchmark for clinical NLP. We evaluate

a wide range of encoder-style LLMs for six different Swedish clinical NLP tasks. This effort represents the first such evaluation to be conducted, and forms a basis for future monitoring of the advances in Swedish clinical NLP.

The results of this first evaluation indicate several interesting trends. The benchmark results suggest that domain adaptation is an effective strategy for improving the performance of LLMs in the clinical domain, at least for small LLMs. Future research should examine whether this also holds for larger models. Furthermore, the evaluations also show that parameter count alone is not enough to perform strongly in the tasks included in our benchmark.

The aim of this paper is to enable monitoring of the progress within Swedish clinical NLP. Due to privacy constraints, the data cannot be shared. We strongly encourage others interested in Swedish clinical NLP to contact us for inclusion in the benchmark. This pragmatic approach to benchmarking enables us to monitor the progress that is being made, which SweClinEval makes possible.

6.1 Limitations

A limitation of the current version of the benchmark is that it only supports encoder models. This is unfortunate, as there is a strong trend towards using autoregressive models both in fine-tuning and few-shot settings. Future versions of the benchmark would benefit from including versions of the datasets that allow non-encoder models to be evaluated. This is not trivial but, as demonstrated by the ScandEval benchmark, it is possible and is an aim for future iterations of the benchmark. Furthermore, we aim to extend the benchmark with more datasets for tasks such as summarization and question-answering.

A more significant limitation of SweClinEval is that currently, only parts of the data can be shared. This restriction is due to privacy regulations surrounding the inherently sensitive clinical data from which the datasets were created. However, two of the datasets – the *Stockholm EPR PHI Corpus* and the *Stockholm EPR ICD-10 Corpus* – are available in automatically de-identified form for academic users. As the regulatory environment around secondary use of private information changes, it may be possible to share the data more freely in the future. For now, our view is that SweClinEval is a pragmatic solution that allows the Swedish NLP community to monitor the progress in Swedish clinical NLP.

References

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. https://doi.org/10.18653/v1/2023.emnlp-main.506 Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. http://arxiv.org/abs/2010.12821 Rethinking embedding coupling in pre-trained language models.

H. Dalianis, A. Henriksson, M. Kvist, S. Velupillai, and R. Weegar. 2015. HEALTH BANK - A workbench for data science applications in healthcare. In *CEUR Workshop Proceedings*. CEUR-WS.

Hercules Dalianis. 2018. https://doi.org/10.1007/978-3-319-78503-5 *Clinical Text Mining*. Springer International Publishing, Cham.

Hercules Dalianis and Sumithra Velupillai. 2010. https://doi.org/10.1186/2041-1480-1-6 De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. *Journal of Biomedical Semantics*, 1(1):6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. https://doi.org/10.18653/v1/N19-1423 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. https://doi.org/10.1016/j.jbi.2023.104286 Dr.bench: Diagnostic reasoning benchmark for clinical natural

language processing. *J. of Biomedical Informatics*, 138(C).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. https://doi.org/10.1145/3458754 Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2022. https://doi.org/10.13026/7VCR-E114 MIMIC-IV.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2021. Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 790–797, Held Online. INCOMA Ltd.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.484 XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

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:1907.11692 [cs]*. ArXiv: 1907.11692.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv:2007.01658 [cs]*. ArXiv: 2007.01658.

Birger Moëll and Fabian Farestam. 2024. https://sltc2024.github.io/abstracts/moell_farestam.pdf Swedish Medical Benchmark, an evaluation framework for LLMs in the Swedish medical domain.

Shawn N. Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C. Chueh, Susanne Churchill, and Isaac Kohane. 2010. https://doi.org/10.1136/jamia.2009.000893 Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association: JAMIA*, 17(2):124–130.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. https://doi.org/10.48550/arXiv.2311.17035 Scalable Extraction of Training Data from (Production) Language Models. ArXiv:2311.17035 [cs].

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Sonja Remmer, Anastasios Lamproudis, and Hercules Dalianis. 2021. Multi-label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT. In *Proceedings of RANLP 2021: Recent Advances in Natural Language Processing, RANLP 2021, 1-3 Sept 2021, Varna, Bulgaria*, pages 1158–1166.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158. Publisher: Elsevier.

Anders Søgaard. 2022. Should We Ban English NLP for a Year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2024. End-to-end pseudonymization of fine-tuned clinical BERT models. *BMC Medical Informatics and Decision Making*, 24:162.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 4245–4252. European Language Resources Association.

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*, volume 30. Curran Associates, Inc.

Sumithra Velupillai. 2011. Automatic classification of factuality levels: A case study on Swedish diagnoses and the impact of local context. In *Fourth International Symposium on Languages in Biology and Medicine, LBM 2011*.

Sumithra Velupillai, Hercules Dalianis, and Maria Kvist. 2011. Factuality levels of diagnoses in Swedish clinical text. *Studies in Health Technology and Informatics*, 169:559–563.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. https://doi.org/10.18653/v1/W18-5446 GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. http://arxiv.org/abs/2402.05672 Multilingual e5 text embeddings: A technical report.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. https://doi.org/10.48550/arXiv.2303.17183 The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling. ArXiv:2303.17183.