Evaluating Text Pre-Processing Strategies for Clinical Document Classification with BERT

1st Sarah Miller, 2nd Serge Sharoff, 2nd Geoffrey Hall
UKRI CDT in AI for Medical Diagnosis and Care
School of Computing, University of Leeds
Leeds, UK
Email:scslmi@leeds.ac.uk, s.sharoff@leeds.ac.uk, g.hall@leeds.ac.uk

3rd Prabhu Arumugam
Multi-Modal Team
Genomics England
London, UK
Email:Prabhu.Arumugam@genomicsengland.co.uk

***Abstract*—In many Natural Language Processing (NLP) tasks, Bidirectional Encoder Representations from Transformers BERT and BERT-based techniques have produced state of the art results. However, this increase in performance comes with a caveat, limitations in the size of the text input the model can process. There are few studies that discuss the constraints of BERTs input length in the context of clinical documents, and as a result, little is known about how effective BERT is in this regard. To overcome these constraints, we investigate techniques for modifying the input text size of pathology report documents. By utilizing various BERT variants, we evaluate these approaches and examine the relative significance of domain specificity versus generic vocabulary training. We demonstrate that BERT models trained on domain knowledge outperform the vocabulary of standard models. In the process of classifying a set of variable-length pathology report texts, BERTs standard truncation approach, which removes text longer than the maximum, performs as well as more sophisticated text pre-processing techniques.**

***Index Terms*—BERT; Clinical Text; Natural Language Processing; Text classification.**

I. Introduction

An essential task that supports clinical workflows throughout health services is information extraction from clinical text documents. Healthcare providers currently invest considerable time and money for clinical specialists to complete this labor-intensive manual task. Automating this process with Natural Language Processing (NLP) has the potential to deliver efficiencies, saving both time and money [1].

Bidirectional Encoder Representations from Transformers (BERT) and BERT-based techniques have shown to deliver notable results across many NLP tasks [2][3]. However, adopting BERT base methods for use with clinical documents presents challenges (1) there is a limit to the input text size the model can process, and (2) they can be computationally demanding, especially during training. BERTs impressive performance can be attributed to its attention mechanism [3][4]. However, what makes BERT so powerful also contributes to its weakness. BERTs attention mechanism scales quadratically and thus limits the size of text input that can be processed by even the most advance computer hardware [5].

Unfortunately, clinical text documents often exceed BERTs maximum input. The maximum input size a BERT model can process is 512 tokens. Tokens are word representations BERT accepts as input, and tokens are not equivalent to words. During the tokenization process words can be split into multiple tokens, therefore, the word count of a document can not be used to determine input size. To address the differences in word to token ratio the inputs into BERT need to be pre-processed or the model architecture needs to be changed to accommodate longer sequences. In this paper, we look to assess the former. Clinical documents are not constrained to a structured format and information included in the texts is at the behest of whomever completes it. Some clinicians are very concise giving all key information in short sentences, whereas some will provide a lengthier description, each approach is clinically valid, but it does present challenges when pre-processing clinical texts [1]. Pre-processing clinical documents with varying formats when there is a limitation on how much of the text can be used is one of those challenges. Key information is distributed throughout documents at varied intervals, and when pre-processing the texts into sections it is difficult to know which sections of the text best contains the text required for classifications.

In our experiments, we evaluate four different text pre-processing strategies to investigate these challenges. We use three variants of BERT models on a multi-label clinical document classification task, using a set of cancer pathology reports from the Genomics England research environment [6]. To the best of our knowledge, there is only one study that investigates the impact of BERTs input size using pathology reports and our study advances their techniques. Our study is also the only study in this area that offers insight into how varying text sequence sizes influences results. The remainder of this article is structured as follows. Section II describes related work on BERT models and the input size limitations for clinical documents. Section III provides an explanation of the dataset and methods used in this study. In Section IV, we present the results of our experiments, and we conclude our findings in Section V.

II. Related Work

Research addressing the input size limitations of BERT has not received much attention in the clinical domain. The automation of ICD coding is the common goal of the few studies in this field and except for one study, all use the MIMIC-III database [7] discharge summaries for their tasks. However,

 ISBN: 978-1-68558-136-7

the results produced across these studies are not entirely consistent. For instance, even after using text pre-processing techniques to overcome BERTs input size constraints in [8][9] the authors discover that simpler networks perform better than BERT. Contrastingly, in [10][11] the authors find that BERT outperforms the simpler models when modifying for input size.

The text pre-processing methods used in these studies follow two approaches (1) truncation (from the right) of any text that exceeds the maximum input size or (2) hierarchical text pre-processing which involves splitting the text into n length segments with or without overlapping. The model individually processes each of the document segments, and to get the classification results for a document in its entirety, each of the segment outputs are combined using either a pooling or attention-based method.

To the best of our knowledge, only one study has investigated how input size restrictions affects other kinds of clinical texts. In [8] the authors use BlueBERT [12] to classify a set of cancer pathology reports as well as the MIMIC-III discharge summaries. They do not use the pathology reports for the ICD coding task. The pathology reports have a set of six document labels, but rather than using a multi-label classification approach, they train six individual models, one for each of the labels. Unlike the results produced for the ICD coding of MIMIC-III discharge summaries, there is no significant difference between BlueBERT, a CNN, and a HiSAN network when classifying the pathology reports. However, the authors in [8] only assess the models trained on the pathology reports using the hierarchical text pre-processing method and a single variant of BERT, BlueBERT, in their experiments.

Outside of the clinical domain there is one in-depth study that explores strategies to adapt BERT for long document classification. In [13] the authors use the standard BERT model to classify several non-clinical datasets and they find that taking the first 128 tokens and last 382 tokens of each document produces the best overall results. In [8] the authors argue this approach may not translate well to the clinical domain but they do not assess this method in any of their experiments. Therefore, in this paper we aim to fill in the gaps between these studies, by systematically investigating how to adapt BERT for the classification of pathology report texts irrespective of their length, and how different variants of BERT perform with the adaptations.

III. Methodology

In this section we present the techniques used in this study. First, we describe the dataset, secondly the models hyperparameters and tokenization settings, and lastly the text pre-processing strategies used for managing longer texts.

A. Dataset

The dataset used in the experiments is a curated dataset taken from the Genomics England research environment. In the dataset there are 15,825 plain text pathology reports for 5413 participants registered on the 100k genome project. The dataset contains reports for participants with three common types of cancer: breast, colorectal, and lung. Classification labels are provided by linking associated clinical records with the date and a tumour id. The dataset is multi-label and multi-class containing a total of 13 classes. The classes in the dataset were transformed into a multi-label set of features to make model training more efficient. Table I displays the dataset features and the distribution. The data is split into a training set of 7753, a validation set of 4748, and a test set of 3324.

B. Models and Hyperprameters

We installed three BERT models from the Huggingface model hub and followed the transfer learning approach. For sequence classification tasks in a multi-label setting, we use a sequence classification instance of the BERT models initialized with pre-trained parameters and fine-tune them for our task. For information on fine-tuning BERT models, we refer readers to resources available in [3][14]. The models used in this study are: (1) BERT-base-uncased [3] implemented as a baseline to compare the performance of the generic BERT vocabulary to clinical ones. (2) Bio_ClinicalBERT [15] which we opted to use because it has been pre-trained using all of PubMed and all MIMIC-III texts, rather than BlueBERT that has been trained with less of the data in both these datasets. (3) BiomedBERT (abstracts + full text) [16] is a model that is pretrained on just PubMed. However, the authors claim it is still superior at biomedical NLP tasks because of its succinct medical vocabulary for tokenization. To perform the document classifications

TABLE I
DATASET FEATURE DISTRIBUTION

Column	**Dataset distribution per label/class label**		
Label	***Features***	***Reports Per Class***	***Total Reports***
Disease Type	Breast	7767	15825
	Colorectal	6389	
	Lung	1668	
Histology Code	80703	985	15825
	81403	6664	
	84803	628	
	85003	6310	
	84803	1238	
Grade	80703	985	15825
	81403	6664	
	84803	628	
	85003	6310	
	84803	1238	

 ISBN: 978-1-68558-136-7

the pathology reports are fed into each of the BERT models, and it is the hidden state h, of the special [CLS] token, produced by BERT, which provides the classification. Because the dataset is multi-label the h[CLS] token, the models document representation, is passed through a sigmoid activation function to produce probabilities for each of the class labels. For further information regarding the [CLS] and other special tokens we refer readers to [3][17]. All three models are trained using an AWS Sagemaker ml.p3.2xlarge instance. Throughout literature training parameters vary for BERT models but for our experiments we opted for 3 epochs, because when we increased this value there was minimal to no difference gained in performance. Likewise for selecting the batch size and learning rate, we found that a batch size of 16 and a learning rate of 3-5e, using an Adam optimizer were the most optimal settings for our task.

C. BERT Tokenization

The BERT tokenizer converts text sequences into word piece tokens. Word piece tokens are words that have been split into segments. For example, the words learning and learned become learn #ing and learn #ed, making each of these words worth 2 tokens, or 4 tokens in total. The word to token ratio given throughout literature is approx., 400 words = 512 tokens and because the word to token limit can only be approximated, we split documents using the token length.

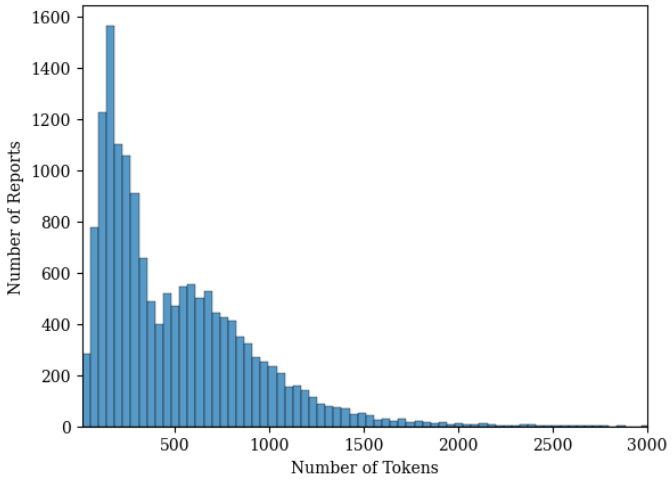

Fig. 1. Distribution of Report Token Lengths

To achieve this the pathology reports are passed through the BERT tokenizer to split the words in each report into their tokenized form. Fig 1 shows the distribution of token lengths for the pathology reports in the dataset. The reports vary in length with the shortest being just 10 tokens and the longest 5372. The mean token length for the dataset is 501, and at least 25% of the reports exceed 700 tokens. If a report is under the maximum it is processed in full. Wherever a report exceeds the maximum it is the count of the tokens that are used to split documents in the text pre-processing strategies.

D. Text Pre-Processing Strategies

a) Right and Left Truncation: An approach for handling sequences longer than 512 tokens is to implement a truncation strategy. BERT tokenizers take parameters for the sequence length and the position for truncation. BERT tokenizers offer either left or right truncation. The default setting is from the right and any tokens exceeding the specified length will be cut off from the right-hand side of the sequence. Likewise, with left truncation anything over the maximum is removed, but in this instance, it is removed from the left, from the beginning rather than the end of the sequence. In our experiments we adopt both approaches to truncation, and we use the maximum sequence length of 512.

b) Left+Right Truncate the Middle: Key information is said to be located at the beginning and end of a document. To investigate this further we follow the approach taken by the authors in [13] and take token segments from the beginning and end of the document, and concatenate them. For any document that exceeds the maximum sequence length of 512 we take the first 128 tokens of the document and the last 382, taking 510 tokens in total, leaving room for BERT special tokens. Any text/tokens in the document that fall in between these values are removed.

c) Hierarchical Text Pre-processing: Hierarchical text pre-processing is where long documents are broken up into segments. In this study any pathology report document exceeding the maximum input length is segmented into n=length/510 tokens. Each segment is prefixed with a [CLS] token and appended with a [SEP] token so they are 512 in length. Each segment is processed by the model following the fine-tuning approach. At the output stage each individual segment has a h[CLS] representation, and we apply mean pooling to combine the h[CLS] representations of all the segments giving a single output and the mean of the probabilities for the whole document.

E. Evaluation Metrics

The most commonly applied metrics in literature for evaluating NLP classification models are Accuracy, F1, and ROC-AUC scores [18]. For example, a popular set of NLP tasks for bench-marking models is GLUE [19] where the majority of tasks use Accuracy and or F1 for evaluation [18]. In the studies we reviewed F1, and ROC-AUC are the metrics reported. There is debate amongst the studies we reviewed which F1 metric is the most relevant, some favor macro F1, and other micro F1 scores for multi-label scenarios. In our experiments, we report micro F1, macro F1, and ROC-AUC in line with current literature for comparison.

IV. Experimental Results

The results in Table II and Table III are used to address three key questions: (1) how well the baseline model with a standard vocabulary compares to the domain trained models. (2) are there differences in performance between the two clinically trained models, and (3) how does text pre-processing to manage input sequence length impact classification performance.

 ISBN: 978-1-68558-136-7

TABLE II
DOCUMENT CLASSIFICATION RESULTS

Model	Model Classification Results			
	Text Strategy	*MicroF1*	*MacroF1*	*ROC-AUC*
BERT-base	Right Truncation	0.84	0.59	0.89
BERT-base	Left Truncation	0.82	0.52	0.88
BERT-base	Left+Right	**0.84**	**0.64**	**0.90**
BERT-base	H Mean Pooling	0.84	0.61	0.89
Bio_CBERT	Right Truncation	0.82	0.52	0.88
Bio_CBERT	Left Truncation	**0.84**	**0.67**	**0.89**
Bio_CBERT	Left+Right	0.84	0.62	0.89
Bio_CBERT	H Mean Pooling	0.84	0.63	0.89
BioMBERT	Right Truncation	0.88	0.69	0.92
BioMBERT	Left Truncation	**0.89**	**0.74**	**0.93**
BioMBERT	Left+Right	0.86	0.67	0.90
BioMBERT	H Mean Pooling	**0.90**	**0.74**	**0.93**

Model names abbreviated e.g., Bio_CBERT = Bio_ClinicalBERT

In respect to the first question, the clinical models do have an increase in performance compared to the BERT-base-uncased model. Confirming that domain specific models can offer an increase in performance when performing clinical NLP tasks. To answer question two there is a difference in performance between Bio_ClinicalBERT and BiomedBERT. This supports the studies claims that the BiomedBERT vocabulary is superior to other clinical variants even when they have been trained with more data. The BiomedBERT tokenizer is said to produce fewer word piece tokens than the other models and they attribute this to why it performs better, suggesting quality over quantity of data for the models training.

To address the final question, the BERT-base-uncased model has a slight increase in performance when using the Left+Right text pre-processing strategy. This reflects the results found by the authors in [13], but it is not reproduced in the results from the clinical models. The clinical models show minor differences but offer a slight increase in performance when using the left truncation and hierarchical mean pooling strategies (referred to as H Mean Pooling in Table II). Some pathology reports contain a summary of the key points of the investigation at the end of the report. Both favored text pre-processing strategies for the clinical models include the end of the document and could attribute to the increase in performance when using those strategies. To further address question three we investigate how the truncation of text has affected results by looking at the results over different document length distributions. Table III shows the results of the classifications across different subsets of document length. In Table III we have split the documents into groups using their original token lengths, prior to truncation, e.g., $\geq$1000 = documents with more than 1k tokens, and $\geq$512 $\leq$1000 are documents that have a token length greater than 512 but less than 1000 etc. We then group them also by the text pre-processing strategy used. What the results in Table III demonstrate is that there is a drop in performance with documents exceeding 1000 tokens. This is as expected, because these documents are subject to the most data loss, +50% of the data in these documents is removed. Longer documents contain key information throughout the length of the text, it is unlikely that it is all contained within the selected section, resulting in lost information required by the classifier. The results in Table III also reveal that there is a drop in performance that occurs for documents with token counts under the maximum limit. When the token counts drop below 250, these much shorter documents contain less information. They are lacking the data required for the successful classification of all the document labels. Thus, the shorter documents are also subject to data loss but in this instance because the clinician has perhaps missed information by being too concise. Changes in performance for texts with the highest and lowest token counts are observed across each of the text pre-processing strategies with BiomedBERT and truncation from the left providing the highest overall scores.

V. Conclusion and Future Work

In this study we have investigated how BERTs limitations in input size influences the classification of plain text pathology report documents. We find that there are performance increases when using a domain specific model for the task, and that not all domain model vocabularies are created equal. Similarly, to the other studies we reviewed the hierarchical text pre-processing approach does offer slightly better performance than the standard truncates from the right method. However, we also observed that for the pathology reports taking just the end of the text, truncation from the left performs just as well, and it is also a much faster method. Whilst our results are not entirely comparable to the results in [8], our models achieved higher macro F1 scores when classifying the pathology reports. Something that this study has highlighted is that the input length of a document is not just a factor when it is significantly longer than the maximum, but also when it is much shorter, and information is thus potentially missing. Pathology reports and other similar clinical texts are variable by nature. There are many factors at play that will dictate the content and length of clinical texts and because there is no current unified format or structure there is no guarantee that all information is recorded adequately. To address variations in the format of pathology reports, adopting a standardised approach could improve data quality for both clinicians and subsequent analyses. However, overall, the BERT models in this study performed well irrespective of the variations.

As previously addressed, currently there are limited studies for clinical document classification with BERT models. The ones that do exist use a limited set of documents from the

 ISBN: 978-1-68558-136-7

TABLE III
MACRO-F1 SCORES FOR CLASSIFICATIONS BY TOKEN LENGTH DISTRIBUTION

Text Pre-processing Strategy	**Macro F1 Scores for Token Length Evaluation**		
+ Token Length Distribution	***BERT-base***	***Bio_ClinicalBERT***	***BiomedBERT***
Right≥1000	0.57	0.52	0.66
Right ≥512 ≤1000	0.60	0.53	0.72
Right ≤512 ≥250	0.60	0.52	0.70
Right ≤250	0.57	0.51	0.68
Left ≥1000	0.51	0.60	0.72
Left ≥512 ≤1000	0.53	0.67	**0.76**
Left ≤12 ≥250	0.52	0.69	**0.77**
Left ≤250	0.51	0.66	0.72
Left+Right≥1000	0.58	0.60	0.62
Left+Right ≥512 ≤1000	0.65	0.63	0.70
Left+Right ≤512 ≥250	0.65	0.62	0.68
Left+Right ≤250	0.62	0.61	0.65

MIMIC-III database, and as discussed by [1] this does not provide a comparable enough view of this task. There needs to be more research using a variety of sources and use cases before the limitations of BERT models for clinical document classification can fully be established.

Future work will look at multi-task learning with BERT models and expanding the feature set of the dataset used in this study. Only a subset of the document features available for classification in the Genomics England research environment was used for this study, and there are potential further analyses, with a wider set of feature labels. BERT models are Deep Learning model architectures that are somewhat of a black box [20] and investigating the models output using explainability methods is also a future direction this research could take.

ACKNOWLEDGMENT

The research in this paper is part of a PhD project funded by the UKRI Center for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care (Project Reference: EP/S024336/1). Secondly, this work has also been supported by the Multi-modal team at Genomics England and we would like to give thanks to all team members for their invaluable knowledge and support, and lastly the research was made possible through access to data and findings in the National Genomic Research Library via the Genomics England Research Environment.

REFERENCES

[1] H. Dong et al., “Automated clinical coding: what, why, and where we are?”, npj Digit. Med., vol. 5, no. 1, pp. 1–8, 2022, doi: 10.1038/s41746-022-00705-7.

[2] I. Chalkidis et al., “An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels,” CoRR, vol. 2010.01653, pp. 7503–7515, 2020, doi: 10.18653/v1/2020.emnlp-main.607.

[3] J. Devlin, M-W. W. Chang, K. Lee, and K. Toutanova, “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), 2019, vol. 1, no. Mlm, pp. 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

[4] A. Vaswani et al., “Attention is all you need”, Proc. 31st Int. Conf. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips, pp. 6000–6010, 2017, doi: 10.5555/3295222.3295349.

[5] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The Long-Document Transformer”, ArXiv, vol. 2004.05150, 2020.

[6] M. Caulfield et al., “National Genomic Research Library.” Genomics England, London, UK, 2020. doi: 10.6084/m9.figshare.4530893.v7.

[7] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” Sci Data, vol. 3, p. 160035, 2016, doi: 10.1038/sdata.2016.35.

[8] S. Gao et al., “Limitations of Transformers on Clinical Text Classification,” IEEE J. Biomed. Heal. Informatics, vol. 25, no. 9, pp. 3596–3607, 2021, doi: 10.1109/JBHI.2021.3062322.

[9] [1] S. Ji, M. Hölttä, and P. Marttinen, “Does the magic of

 ISBN: 978-1-68558-136-7

BERT apply to medical code assignment? A quantitative study", Comput. Biol. Med., vol. 139, pp. 1–13, 2021, doi: 10.1016/j.compbiomed.2021.104998.

[10] C. W. Huang, S. C. Tsai, and Y. N. Chen, "PLM-ICD: Automatic ICD Coding with Pretrained Language Models", Clin. 2022 - 4th Work. Clin. Nat. Lang. Process. Proc., pp. 10–20, 2022, doi: 10.18653/v1/2022.clinicalnlp-1.2.

[11] A. Afkanpour et al., "BERT for Long Documents: A Case Study of Automated ICD Coding," LOUHI 2022 - 13th Int. Work. Heal. Text Min. Inf. Anal. Proc. Work., pp. 100–107, 2022.

[12] Y. Peng, S. Yan, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets", in Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, no. iv, pp. 58–65. doi: 10.18653/v1/W19-5006.

[13] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in Chinese Computational Linguistics, 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.

[14] HuggingFace, "Fine-tune a pretrained model", www.huggingface.com. https://huggingface.co/docs/transformers/en/training (accessed Feb. 05, 2024).

[15] E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings", in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78. doi: 10.18653/v1/W19-1909.

[16] Y. U. Gu et al., "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," ACM Trans. Comput. Healthc., vol. 3, no. 1, pp. 1–23, 2020, doi: 10.1145/3458754.

[17] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works", Trans. Assoc. Comput. Linguist., vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.

[18] P. Vickers, L. Barrault, E. Monti, and N. Aletras, "We Need to Talk About Classification Evaluation Metrics in NLP," in Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Nov. 2023, vol. 1, pp. 498–510. [Online]. Available: http://arxiv.org/abs/2401.03831 pp.498-510.

[19] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "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, 2018, vol. 38, no. 3, pp. 353–355. doi: 10.18653/v1/W18-5446.

[20] G. Prasad, Y. Nie, M. Bansal, R. Jia, D. Kiela, and A. Williams, "To what extent do human explanations of model behavior align with actual model behavior?," in BlackboxNLP 2021 - Proceedings of the 4th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2021, pp. 1–14. doi: 10.18653/v1/2021.blackboxnlp-1.1.

 ISBN: 978-1-68558-136-7