



# InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline

Kai Kugler<sup>1</sup>   
Simon Munker<sup>1</sup>   
Johannes Höhmann<sup>1</sup>  
Achim Rettinger<sup>1</sup> 

1. Computational Linguistics & Digital Humanities, University of Trier , Trier, Germany.

## Citation

Kai Kugler, Simon Munker, Johannes Höhmann, and Achim Rettinger (2023). "InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline". In: *Journal of Computational Literary Studies* 2 (1). 10.48694/jcls.3572

**Date published** 2023-03-05

**Date accepted** 2024-02-19

**Date received** 2023-01-10

## Keywords

contextualized word embeddings, derived text formats, text reconstruction, transformer encoder, publication restrictions

## License

CC BY 4.0 

## Reviewers

David Mimno, Mike Kestemont

## Note

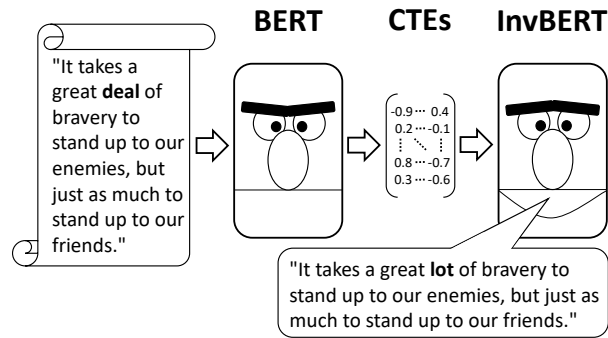
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 2nd Annual Conference of Computational Literary Studies 2023 at Würzburg University.

**Abstract.** Digital Humanities and Computational Literary Studies apply automated methods to enable research on large corpora which are not feasible by manual inspection alone. However, due to copyright restrictions, the availability of relevant digitized literary works is limited. Derived Text Formats (DTFs) have been proposed as a solution. Here, textual materials are transformed in such a way that copyright-critical features are removed, but that the use of certain analytical methods remains possible. Contextualized word embeddings produced by transformer-encoders are promising candidates for DTFs because they allow for state-of-the-art performance on analytical tasks. However, in this paper we demonstrate that under certain conditions, the reconstruction of the original text from token representations becomes feasible. Our attempts to invert BERT suggest that publishing the encoder together with the contextualized embeddings is unsafe, since it allows to generate data to train a decoder with a reconstruction accuracy sufficient to violate copyright laws.

## 1. Introduction

Due to copyright laws the availability of more recent text material (specifically literary works from the last 100 years) for scientific analyses is quite limited. For disciplines such as Computational Literary Studies (CLS), these legal restrictions make research on contemporary literature difficult because the relevant primary texts may not be published with the research results (e.g. to enable follow-up research), as current principles of scientific data management demand (Wilkinson et al. 2016). Depending on national law, there might be some degree of freedom to use protected texts for scientific studies and give reviewers access to them, but in most cases they still can't be published fully, making it hard for the research community to reproduce or build on scientific findings. According to German copyright law, for example, there are now possibilities for making copyright-protected material accessible in the context of scientific research, e.g. for the peer review process, but these exceptions are so narrowly defined that the corpora are no longer available for follow-up research.<sup>1</sup>

1. See § 60d UrhG, [https://www.gesetze-im-internet.de/urhg/\\_\\_60d.html](https://www.gesetze-im-internet.de/urhg/__60d.html).



**Figure 1:** Sample text reconstruction to a Harry Potter quote from Joan K. Rowling (1998) by inverting BERT.

This is a fundamental issue for research in Digital Humanities (DH) and Computational Literary Studies (CLS), but applies also to any analysis of text documents that cannot be made available due to privacy reasons, copyright restrictions or business interests. This, for instance, makes it hard for digital libraries to offer their core service, which is the best possible access to their content. While they provide creative solutions as a compromise, like *data capsules* or *web-based analysis tools*,<sup>2</sup> such access is always limited and complicates subsequent use and reproducibility.

As a consequence, there have been attempts to find a representation formalism which retains as much linguistic information as possible while not disclosing the original text fully. Such text representations have been referred to as Derived Text Formats (DTFs) (Schöch et al. 2020a). While such DTFs are always a compromise between the degree of obfuscation (non-reconstructibility) and degree of analyzability (retained information), there are DTFs with clear advantages over others. In the end, they should always be more informative than not publishing the documents at all.

We investigate whether Contextualized Token Embeddings (CTE), like the ones obtained from a transformer encoder stack trained on a self-supervised masked language modeling (MLM) task (Devlin et al. 2019), are a promising candidate for DTFs. On the one hand, they are the state-of-the-art text representation for most Natural Language Understanding tasks (Wang et al. 2019a,b), including tasks relevant to DH and CLS, like text classification, sentiment analysis, authorship attribution or text re-use (Schöch et al. 2020b). On the other hand, it appears difficult to reconstruct the original text just from its CTEs because, unlike (static) word embeddings, there is no fixed inventory of representations that does not change from sentence to sentence. Thus, we pose the following research question:

In which scenarios can protected text documents be released publicly if they are encoded as contextualized embeddings without the original content being able to be reconstructed to an extent that the publication violates copyright laws?

After presenting related work (section 2) we will first formalize different reconstruction scenarios, which allow us to define potential lines of attack that aim at reconstructing the original text (section 3). Next, we will discuss the feasibility of each line of attack.

2. See [https://www.hathitrust.org/htrc\\_access\\_use](https://www.hathitrust.org/htrc_access_use).

In [section 4](#) we focus on the most promising lines of attack by evaluating their feasibility empirically ([section 5](#)), before concluding in [section 6](#).

## 2. Related Work

First, we look at the very recent field of DTFs, before presenting existing work on text reconstruction beyond copyright protected texts.

### 2.1 Derived Text Formats

DTFs, like n-grams or term-document matrices, are an important tool to the Digital Humanities and Computational Linguistics, since they allow the application of quantitative methods to their research objects. However, they have another important advantage: If the publication of an original text is prohibited, DTFs may still enable reproducibility of research (Schöch et al. [2020a,b](#)). This is especially important for CLS, where there is only a small “window of opportunity” of available texts from the year 1800 to 1920 due to technical issues on the lower and copyright restrictions on the upper boundary. Since this is of permanent concern and an obstacle to open science, tools to widen this window are of great importance to the field. Other approaches to tackle this issue, like granting access to protected texts in a closed room setting, come with their own major drawbacks and still do not enable an unhindered exchange of scientific findings. Therefore, in most cases, DTFs like term-document matrices are the best solution available. The aim of these formats is to retain as much information as possible while minimizing reconstructibility. In reality, however, the latter most often is achieved by compromising on the former. This leads to the variety of feasible analytical down-stream tasks being narrowed. A format that preserves a noticeable amount of information and is already used as a DTF are word embeddings like Word2Vec (Mikolov et al. [2013](#)) or GloVe (Pennington et al. [2014](#)). However, similar to term-document matrices, they can only be applied to document-level tasks. Otherwise, there remains considerable doubt regarding their resilience against reconstruction attempts. A promising attempt to address this problem is by using contextualized word- or more precisely token-embeddings (CTEs) generated by pretrained language models instead, since the search space for identifying a token grows exponentially with the length of the sequence containing it. Additionally, these embeddings carry even more, especially lexical semantic information (Vulic et al. [2020](#)) and achieve state-of-the-art results on various down-stream tasks.

### 2.2 Reconstruction of Information from Contextualized Embeddings

Recently, attention was drawn to privacy and security concerns regarding large language models by prominent voices in ethics in AI (Bender et al. [2021](#)), as well as a collaborative publication of the industry giants Google, OpenAI and Apple (Carlini et al. [2021](#)). In the latter, the authors demonstrated that these models memorize training data to such an extent that it is not only possible to test whether the training data contained a given sequence (membership inference; Shokri et al. [2017](#)), but also to directly query samples from it (training data extraction). Other recent research supports these findings and agrees that this problem is not simply caused by overfitting (Song and Shmatikov [2019](#); Thomas et al. [2020](#)). Large language models like GPT-3 (Brown et al. [2020](#)) or T5 (Raffel

et al. 2020) were trained on almost the entirety of the available web, which poses a special concern, since sensitive information like social security numbers is unintentionally being included. Hence, a majority of the literature focuses on retrieving information about the training data. However, we argue that such attacks are less successful in the case of literary works, since (a) the goal in this scenario would usually be the reconstruction of a specific work, and (b) the attacks are not suited to recover more than isolated sequences.

A third prominent type of attack which can be performed quite effectively and reveals some information about training data is attribute inference (Mahloujifar et al. 2021; Melis et al. 2019; Song and Raghunathan 2020). It is also of little relevance, since it aims to infer information like authorship from the embeddings, which is non-confidential in a DTF setting anyway. More so, authorship attribution is actually a relevant field of research in the DH.

The main threat regarding CTEs as DTFs are embedding inversion attacks, where the goal is the reconstruction of the original textual work they represent. However, research on this topic is still limited and most papers focus on privacy rather than copyright. Therefore, very few go beyond the retrieval of isolated sensitive information. For example, Pan et al. (2020) showed that it is possible to use pattern-recognition and keyword-inference techniques to identify content with fixed format (e.g. birth dates) or specific keywords (e.g. disease sites) with varying degree of success (up to 62% and above 75% average precision, respectively). However, this is easier and the search space is smaller than in the case of reconstructing full sequences drawn from the whole vocabulary.

To the best of our knowledge, retrieval of the full original text is covered only by Song and Raghunathan (2020). Using an RNN with multi-set prediction loss in a setting with access to the encoding model as a black-box, they were able to achieve an in-domain F1 score of 59.76 on BERT embeddings. However, since privacy was their concern, they did not consider word ordering in their evaluation, which is crucial when dealing with literary works. Therefore, and since they failed to improve on their results using a white-box approach as well, we believe that the security of the usage of CTEs as DTFs still remains an unanswered question.

When dealing with partial-white- or black-box scenarios, a final type of attack should be kept in mind: Inferences about the model itself. Even though not the goal here, successful model extraction attacks (Krishna et al. 2020) may transform a black-box situation into a white-box case. However, critical information can even be revealed by fairly easy procedures like model fingerprinting. This was showcased on eight state-of-the-art models by Song and Raghunathan (2020), who were able to identify the model based on a respective embedding with 100% accuracy.

### 3. Reconstruction Task and Attack Vectors

This paper is not about improving or applying transformers, but about inverting them. To introduce reconstruction models (Rigaki and Garcia 2020), we first describe scenarios for possible attacks. Then, we lay out different attack vectors based on the scenarios.

### 3.1 Reconstruction Scenarios

Formally, the reconstruction scenarios can be defined as follows:

**Given:** Contextualized token embeddings (CTEs) of a copyright protected literary document  $W$  (typically a book, containing literary works, like poetry, prose or drama) are made available in every scenario. Depending on the scenario, additional information is available:

**WB - White Box Scenario:** The most flexible scenario is given if the encoder  $enc()$ , including the neural network's architecture and learned parameters, and the tokenizer  $tok()$  are made openly available in addition to the CTEs. In this case, analytical experiments can be conducted by DH researchers that require to adapt/optimize the encoder  $enc()$  and/or the tokenizer  $tok()$ .

**BB - Black Box Scenario:** A scenario with little flexibility from the perspective of a DH researcher is given when the tokenizer  $tok()$  and the encoder  $enc()$  are made available as one single opaque function and are only accessible for generating mappings from  $W$  to CTE. A similar scenario arises if ground truth training data is available (i.e., aligned pairs of  $W$ s to CTEs are given). In this case, the researchers are still able to label their own training data and use it to optimize  $enc()$  or embed other data not yet available as CTEs for analysis. However, if provided as a service, the number of queries allowed to be sent to  $enc()$  might be limited up to a point where the model is not released at all. Then, existing implementations can be reused in order to perform a standard analytical task if the respective task-specific top layer function is also provided. Note that a BB can be turned into WB by successful model extraction attacks.

**GB - Gray Box Scenario:** If the encoder-transformer pipeline with  $tok()$  and  $enc()$  used for generating CTEs is available to some degree (e.g., the tokenizer is given), we refer to it as a Gray Box (GB) scenario.

**Searched:** A function or algorithm  $inv(CTE) = \hat{W}$  that inverts the model pipeline or approximates its inverse and outputs reconstructed text  $\hat{W}$  from CTEs.

### 3.2 Inversion Attacks

We consider three lines of attack:

**Inverting Functions:** Inverting  $enc()$  and  $tok()$  using calculus requires an attacker to find a closed-form expression for  $tok^{-1}()$  and  $enc^{-1}()$ . Since this requires knowledge of the parameters of the encoder pipeline, this is only applicable to a WB scenario. Even then, this approach would only be feasible if all functions in question are invertible, which is not the case for BERT-like transformer-encoder stacks.

**Exhaustive Search:** Sentence-by-sentence combinatorial testing of automatically generated input sequences to "guess" the contextualized token embeddings would be applicable to WB, GB and BB, as long as an unlimited number of queries to  $enc()$  is allowed. However, combinatorial explosion renders this approach infeasible: A sentence of 15 tokens results in  $18 \cdot 10^{66}$  possible combinations, assuming a vocabulary size of 30,522 different tokens, like in the case of BERT<sub>BASE</sub>.

**Machine Learning:** Learning an approximation of  $tok^{-1}(enc^{-1}())$  can be attempted as soon as training samples are available or can be generated. We assume that an attack is more likely to be successful if components of the embedding generating pipeline are accessible, because in a GB scenario the components can be estimated separately, reducing the complexity compared to an end-to-end BB scenario.

Since a successful BB attack works equally well in a GB scenario and a successful GB attack works in a WB scenario, we restrict our empirical investigation to two machine learning based attacks, one for a GB, where  $tok()$  is given, and one for the BB scenario. We call our GB attack *InvBert Classify* and our BB attack *InvBert Seq2Seq*. Both models are detailed in [Figure 2](#) and described in the next section.

## 4. Experimental Design

In this section, we describe two attack models, one for a GB and one for a BB scenario, introduced in [section 3](#). First, we introduce and discuss the datasets. Next, we explain both neural network structures and the general attack pipeline. The code and datasets are publicly available in a Github repository (see [section 7](#) and [section 8](#)).

### 4.1 Data

As a data basis, we have chosen two text corpora that fulfil three conditions: First, the texts of the corpora should be similar to the protected works that are to be distributed in DTF. We restrict ourselves to English-language prose texts and choose the corpora accordingly. Secondly, the corpora must be big enough to draw datasets of a size that allow models to be trained successfully. Furthermore, it is important to us that our results are reproducible, which is why we have chosen openly available data.

First, we scraped the *Archive of Our Own (AO3)*, an openly available fanfiction repository, using a modified version of *AO3Scrapper*.<sup>3</sup> During the preprocessing step, we filtered out mature, extreme, and non-general audience content using the available tags. We split the AO3 data into the following three topics (based on the ten most common tags *Action*, *Drama* and *Fluff*<sup>4</sup>) to get different samples. [Table 1](#) shows the exact size and number of samples of each subset.

Before training our models, the datasets were each split into non-overlapping training and evaluation datasets. Training is performed on the complete training dataset (100%) or on a subset of this data (10%, 1%, 0.1%) to assess how the amount of training data affects the text reconstruction ability of the models.

As fanfiction mostly resembles contemporary literature, we gathered a fourth dataset from Project Gutenberg, a non-commercial platform with a focus on archiving and distributing historical literature, including western novels, poetry, short stories, and drama.<sup>5</sup> Consequently, our Gutenberg training / evaluation dataset contains a mix of different genres. This dataset shows a slightly higher Type-Token-Ratio, indicating a

3. See <https://archiveofourown.org> and <https://github.com/radiolarian/AO3Scrapper>.

4. "Feel good" fan fiction designed to be happy, and nothing else, according to [https://en.wikipedia.org/wiki/Fan\\_fiction](https://en.wikipedia.org/wiki/Fan_fiction).

5. See <https://www.gutenberg.org/>.

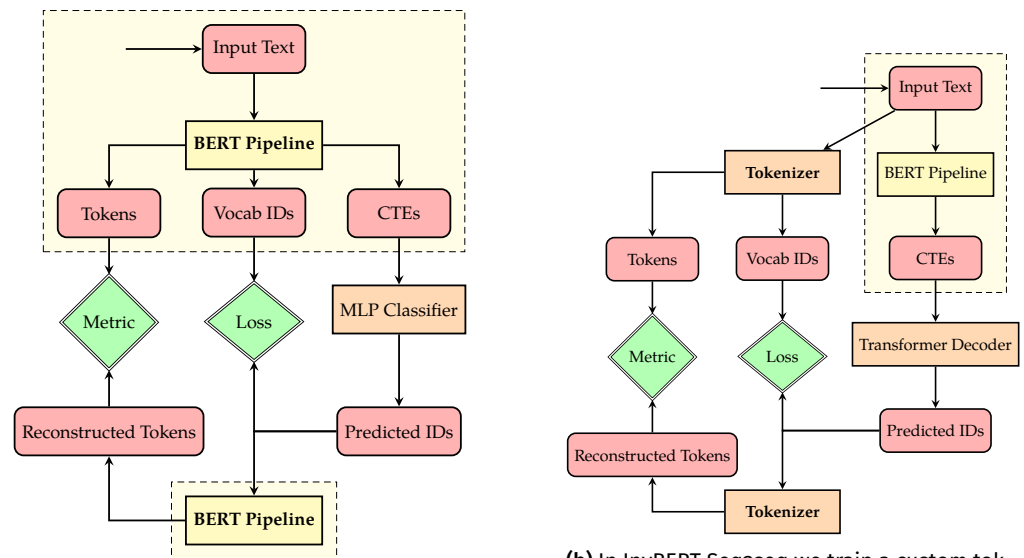
Name	Filesize	Tokens	Unique	Ratio
Action	372.97 MB	72,086,159	152,847	0.002120
Drama	304.51 MB	59,133,691	136,759	0.002313
Fluff	327.80 MB	64,002,888	144,492	0.002258
Gutenberg	257.94 MB	48,807,783	173,716	0.003559

**Table 1:** Size and number of contained training samples of the collected data sets, number of unique tokens (types) and Type-Token-Ratio.

higher lexical variation in contrast to the AO<sub>3</sub> datasets (see Table 1). Gutenberg’s content is sorted by *bookshelves*; we have selected prose genres in Modern English (Classics, Fiction, Adventure etc.) and have not been removing any metadata.

## 4.2 Models & Pipelines

In section 3, we argued that machine learning models are promising candidates for inversion attacks. We propose two models, one for a GB and one for a BB scenario:



(a) Our InvBERT Classify approach retrieves tokens, IDs, CTEs from the encoder (BERT Pipeline) and utilizes a multi-layer classifier to predict IDs. We use the identical encoder to reconstruct the original token/text.

(b) In InvBERT Seq2seq we train a custom tokenizer (BytePair) and utilize only the given CTEs (BERT Pipeline) to sequentially predict token IDs utilizing a Transformer Decoder Structure. Here we use our tokenizer to reconstruct the original token/text.

**Figure 2:** Flowchart for each approach. Givens are enclosed in a dotted yellow area and attack-specific modules to be estimated are filled with orange. Data objects are highlighted in red, while green represent the evaluation/objective function.

**InvBERT Classify (GB):** Here, we have access to the CTEs and the tokenizer  $tok()$ . As the tokenizer is a look-up table that can be queried from both directions, the inverse  $tok^{-1}()$  to  $tok()$  is also provided, effectively simplifying the problem of finding an approximation of the inverse  $tok^{-1}(enc^{-1}())$  of the whole pipeline to just  $enc^{-1}()$ . We train a multi-layer perceptron to predict the vocabulary IDs given CTEs. As we use the given tokenizer, CTEs and IDs have a one-to-one mapping, and our attack boils down to a high-dimensional token classification task.

**InvBERT Seq2Seq (BB):** Here, we only have access to the CTEs. Without the tokenizer, we lose the one-to-one mapping and cannot infer the token CTE ratio. Thus, we have to train a custom tokenizer and optimize a transformer decoder structure to predict our sequence of custom input IDs. The decoder utilizes complete sentence CTEs as generator memory and predicts each token ID sequentially.

We use the *Hugging Face API*<sup>6</sup> to construct a batch-enabled BERT Pipeline capable of encoding plain text into CTEs and decoding (sub-)token IDs into words. All parameters inside the pipeline are disabled for gradient optimization. Our models and the training/evaluation routine are based on *PyTorch modules*.<sup>7</sup> We utilize AdamW as an optimizer and the basic cross-entropy loss. Our model implementations have ~ 24M (InvBert Classify) and ~ 93M (InvBERT Seq2Seq) trainable parameters. We train on a single Tesla V100-PCIe-32GB GPU and do not perform any hyperparameter optimization. Further, we use in each type of attack the identical hyperparameter settings to ensure the highest possible comparability.<sup>8</sup> A training epoch for a model takes up to 8 hours, depending on the dataset and type of attack.

### 4.3 Evaluation Metrics

We evaluate the 3-gram, 4-gram, and sentence precision in addition to the BLEU metric (Papineni et al. 2002). The objective of our model is to reconstruct the given input as closely as possible. BLEU defines our lower bound in terms of precision, as it is based on n-gram precision allowing inaccurate sentences with matching sub-sequences. Since the BLEU metric might be too imprecise to quantify whether a reconstruction captures the content of a sentence and style of the author, we preferred to use complete sentence accuracy in our quantitative evaluation. There, we only count perfectly correct reconstructions, resulting in a significantly higher bound in contrast to BLEU. While the BLEU metric can give us an indication of how closely the reconstruction candidate resembles the original text, we consider correct reconstructions to be a clear sign that possible copyright violations are imminent when publishing.

## 5. Empirical Results

In this section we will first present our qualitative results, before showing some examples of different reconstruction results.

### 5.1 Quantitative Evaluation

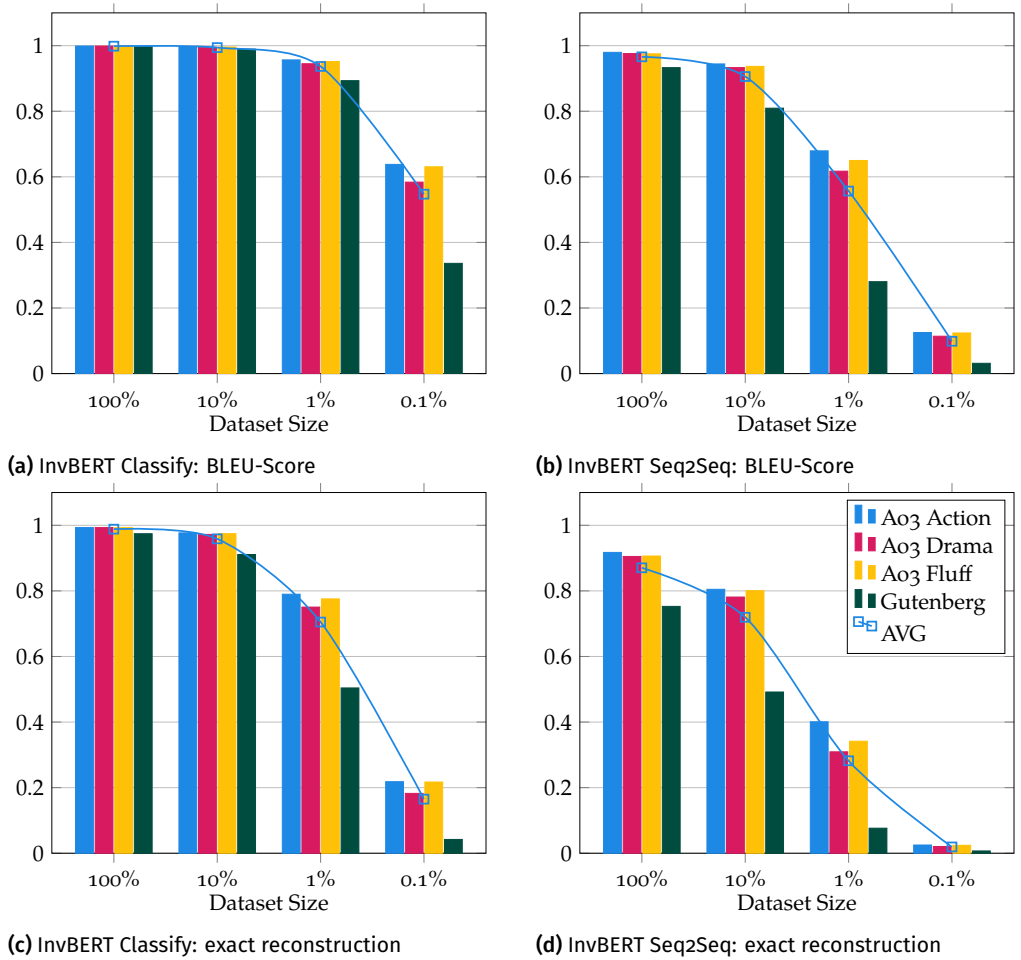
We quantitatively evaluated the trained models in-domain by calculating their sentence accuracy over all samples of their corresponding test set. Additionally, we conducted the reconstruction across all other evaluation datasets to measure the out-of-domain performance. A condensed representation of our in-domain results is presented in [Figure 3](#), while the full results are included in [Appendix A](#).

6. See <https://huggingface.co>.

7. See <https://pytorch.org>.

8. The parameters used for the experiments can be found in the configuration files of the repository.





**Figure 3:** Both reconstruction approaches compared by their in-domain BLEU score (a), (b) and exact sentence reconstruction accuracy (c), (d) on the evaluation data sets.

The InvBERT Classify model achieves a very high in-domain as well as out-domain sentence reconstruction accuracy when trained on 100% and 10% of the training dataset. Thus, we can reconstruct around  $\approx 97\%$  of the original content without errors. Even when just utilizing 1% of the training datasets, our model scores  $\approx 65\%$  sentence reconstruction accuracy. This likely still is enough to violate copyright laws since the remaining 35% of sentences get very close to the originals. In comparison, the sentences generated with a model trained on only 0.1% of the data no longer resemble the original input data.

The consistently high BLEU-scores achieved by both models, even with smaller datasets (BLEU-scores  $> 60\%$  even if only 1% of the data is used), indicate that the text reconstructions are very close to the original text and that perhaps just individual tokens could not be reconstructed exactly.

We observe that the performance on the samples of the AO3 dataset is very consistent. The performance considerably drops on the Gutenberg corpus. We assume that the more heterogeneous content in combination with input shuffling during training yields a more challenging dataset than our AO3 crawl. In particular, the smaller the training subsets, the smaller the number of samples of a certain genre inside our Gutenberg corpus. Additionally, the Gutenberg corpus contains noise like metadata and unique tokens in the form of title pages and table of contents, which we did not clean. The differences are negligible when using 100% or 10% of the training dataset, but become clear on 1% or 0.1% training data usage, where the accuracy differs by around 20%.

The InvBERT Seq2Seq2 model reaches slightly worse results while also being much more sensitive to the training data size and the type of dataset. This is to be expected, since this approach utilizes a more complex network architecture that sequentially predicts the reconstruction parts. We attribute the differences to the more complex task and the higher number of trainable parameters.

## 5.2 Qualitative Evaluation

To put our previously made assumption about their reconstruction quality to the test, we applied our models to 15 quotes from the Harry Potter book series.<sup>9</sup> The calculated metrics in Table 2 show that the performance on these real-world examples are consistent with the quantitative results on our test data.<sup>10</sup>

InvBERT Classify completely reconstructs the samples when trained on 100% or 10% of the training dataset. Only when using 1% or 0.1% of the training data, the model predicts false but semantically similar content. By contrast, InvBERT Seq2Seq starts to produce substantial errors in its reconstruction while using 10% of the training data, and with less data, the predictions do not resemble a reasonable reconstruction attempt neither on the syntactic nor on the semantic level.

9. Retrieved from <https://mashable.com/article/best-harry-potter-quotes>.

10. Reconstructions of the 15 quotes by all 32 models trained on the different datasets of different sizes can be found in the repository provided; see the respective logfiles of the models.

<b>SRC:</b>	if you want to know what a man's like, take a good look at how he treats his inferiors, not his equals.	i'll just go down and have some pudding and wait for it all to turn up ... it always does in the end.
<b>InvBERT Classify</b>		
100%	<i>exact reconstruction</i>	<i>exact reconstruction</i>
10%	<i>exact reconstruction</i>	<i>exact reconstruction</i>
1%	if you want to know what a man's like, take a good look at how he treats his <b>subordinates</b> , not his equals.	<i>exact reconstruction</i>
0.1%	if you want to know what a man's like, take a good look at how he treat his <b>enemies</b> , not his <b>friends</b> .	i'll just go down and have some <b>dinner</b> and wait for it all to come up ... it always does in the end.
<b>InvBERT Seq2seq</b>		
100%	<i>exact reconstruction</i>	<i>exact reconstruction</i>
10%	if you want to know what a man's like, take a good look at how he treats his inferior <b>tors</b> , not his equals.	<i>exact reconstruction</i>
1%	if you want to know what a man's like, take a good look at <b>his partners</b> , not <b>his partners</b> .	i'll just go down and <b>wait for</b> some <b>chocolate</b> and wait for it all to turn up <b>in the end</b> ... it always does in the end.
0.1%	if you <b>have a little to get a little, but you're a little look at him, not like he're a little look.</b>	<b>i'll go and get up up the rest of the rest , it just just just have been going to get up.</b>

**Table 2:** Example of Harry Potter quotes J. K. Rowling 2006 and their predictions. Differences are highlighted: **red** as error and **yellow** as false, but semantically acceptable. 'exact reconstruction' represents identical reproduction.

### 5.3 Discussion

Our exemplary manual evaluation corroborates the results from our quantitative experiments. Both attacks can, if enough data is available, successfully reconstruct the original content. In conclusion, according to our assessment, all scenarios (WB, GB and BB) cannot be considered safe. Even in the "safest" BB scenario without a given tokenizer, reconstruction is feasible.

Collecting training data has proven to be very easy, as there are many corpora available digitally that are sufficiently similar to modern English-language texts. The word order information that BERT can extract from this data is apparently sufficient to reconstruct texts from CTEs derived from texts that are not allowed to be published.

Thus, copyright violations are imminent when publishing CTEs as DTFs.

## 6. Conclusion and Future Work

To conclude, we first summarize our contributions and findings, before outlining open research questions.

## 6.1 Summary and Conclusion

Derived Text Formats (DTFs) are an important topic in Digital Humanities (DH). In this field, the proposed DTFs rely on deleting important information from the text, e.g., by using term-document matrices or paragraph-wise randomising of word order. In contrast, Contextualized Token Embeddings (CTEs), as produced by modern language models, are superior in retaining syntactic and semantic information of the original documents. However, the use of CTEs for large-scale publishing of copyright-protected works as DTFs is constrained by the risk that the original texts can be reconstructed.

In this paper, we first identified and described typical scenarios in DH where analyzing text using CTEs is helpful to different degrees. Next, we listed potential attacks to recover the original texts. We theoretically and empirically investigated what attack can be applied in which scenario.

Our findings suggest that if a certain number of training instances (known mappings of sequences of CTEs produced by the encoder to the original sentences) are given or can be obtained, it is not safe to publish CTEs. Even the safest BB scenario that we covered in this paper is not resistant against reconstruction attacks. Consequently, all GB and WB scenarios are even more vulnerable.

## 6.2 Future Work

While researchers in DH have to judge the usefulness of CTEs as DTFs, finding a copyright-compliant way of publishing content is also relevant for the field of Natural Language Processing (NLP) in general. In this field, CTEs have only been investigated with respect to privacy risks, but not copyright protection. After all, the problem of reproducibility of scientific results from restricted corpora is not limited to the DH. Therefore we encourage the establishment of a novel research niche. The focus of this paper is to define the task of reconstructing text from CTEs of literary works. Accordingly, we only covered the most obvious lines of attack; there are more scenarios that require additional investigation.

Another potential scenario that has not been discussed in this paper is the publication of CTEs without any (means to generate) training data. Although this scenario seems conceivable, there are practical reasons that virtually rule it out: First, to be of any value for DH researchers, the bibliographic metadata (author, title, ...) about the literary work has to be published along with the CTEs. In addition, the rich information encoded in CTEs (e.g., compared to a bag-of-words representation) is more likely to be useful when used in conjunction with more detailed information such as sentence boundaries. Second, ensuring that no training data can be obtained from a released sequence of CTEs seems only feasible in very special cases. If (parts of) the literary works in the corpus can be obtained in a digitized format through other means, it might be possible to align them with the sequence of CTEs and generate a training set. How sentences can be aligned remains the key research challenge in such a scenario, but as soon as an alignment can be established, it becomes an invertible BB scenario.

Also, there is the question of finding a compromise scenario where the complete sequence of CTEs is not published or noise is added, as has been done with DTFs. Examples

are shuffling the sequence, random deletion of a portion of the CTEs, or representation of certain CTEs by linguistic features. What benefits CTEs provide in such scenarios is also a question for future research.

While we covered the most obvious lines of attack in this paper, there are more scenarios that require additional investigation: Potential combinations of different DTFs or metadata might allow new lines of attack, for instance, if n-grams plus CTEs are published for the same text. Moreover, a mapping between the used embedding and a different embedding, based on the incorporated linguistic information they share, might be possible.

CTEs generated by more modern language models than BERT are also of interest for future research. These models keep growing in size and capabilities, as does the complexity of the CTEs they generate. It is to be investigated whether texts represented by these embeddings can still be reconstructed using approaches like ours, but we assume that it is a matter of scaling the reconstruction model accordingly, rather than rendering our general approach infeasible.

Opposite to the attack perspective, an open research question is whether there are novel types of DTFs, beyond CTEs, which are more expressive and more safe.

Another related issue that we did not discuss is the suitability of quantitative metrics for measuring copyright violations. Ultimately, it is a legal consideration, if a reconstruction accuracy, e.g., above a certain BLEU-score, violates copyright laws. This is beyond the scope of this paper.

Ultimately, publishers and libraries need to decide whether they release DTFs of their inventory. However, based on our findings, we advise against it, since it is likely that training samples might be obtained. Still, we believe that more research is needed to find compromise solutions that balance usefulness while ensuring safety from reconstruction. What contribution CTEs can provide is still an open question.

For researchers, this is an exciting challenge, since it requires both theoretical studies regarding computational complexity and empirical experiments with real-world corpora in real-world settings.

## 7. Data Availability

The AO<sub>3</sub> corpus cannot be made available, for the same copyright reasons discussed in this paper. However, it can be recrawled to replicate our experiments. The Gutenberg corpus is freely downloadable and usable: <https://gitlab.rlp.net/cl-trier/InvBERT>.

## 8. Software Availability

The code to replicate our findings is available on GitHub, once the paper is accepted (during review as an anonymous repository): <https://gitlab.rlp.net/cl-trier/InvBERT>.

## 9. Author Contributions

**Kai Kugler:** Conceptualization, Data curation, Validation, Writing

**Simon Münker:** Methodology, Software, Formal Analysis, Visualization, Writing

**Johannes Höhmann:** Methodology, Visualization, Writing

**Achim Rettinger:** Conceptualization, Project administration, Supervision, Writing

## References

- Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 10.1145/3442188.3445922.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. Vol. 33. Curran Associates, Inc., 1877–1901. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf).
- Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel (2021). “Extracting Training Data from Large Language Models”. In: *USENIX Security Symposium*. <https://www.usenix.org/system/files/sec21-carlini-extracting.pdf>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “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)*. 10.18653/v1/N19-1423.
- Krishna, Kalpesh, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer (2020). “Thieves on Sesame Street! Model Extraction of BERT-based APIs”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. <https://doi.org/10.48550/arXiv.1910.12366>.
- Mahloujifar, Saeed, Huseyin A. Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa (2021). “Membership Inference on Word Embedding and Beyond”. In: *ArXiv*. <https://doi.org/10.48550/arXiv.2106.11384>.
- Melis, Luca, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov (2019). “Exploiting unintended feature leakage in collaborative learning”. In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 10.1109/SP.2019.00029.

- Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. [10.48550/arXiv.1301.3781](https://arxiv.org/abs/1301.3781).
- Pan, Xudong, Mi Zhang, Shouling Ji, and Min Yang (2020). “Privacy risks of general-purpose language models”. In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. [10.1109/SP40000.2020.00095](https://doi.org/10.1109/SP40000.2020.00095).
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135).
- Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162).
- Raffel, Colin, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *ArXiv*. [10.48550/arXiv.1910.10683](https://arxiv.org/abs/1910.10683).
- Rigaki, Maria and Sebastian Garcia (2020). “A Survey of Privacy Attacks in Machine Learning”. In: *arXiv*. [10.48550/arXiv.2007.07646](https://arxiv.org/abs/2007.07646).
- Rowling, J. K. (2006). *Harry Potter and the Half-Blood Prince*. Bloomsbury.
- Rowling, Joan K. (1998). *Harry Potter and the Sorcerer’s Stone*.
- Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020a). “Abgeleitete Textformate: Prinzip und Beispiele”. In: *RuZ-Recht und Zugang* 1.2. [10.5771/2699-1284-2020-2-160](https://doi.org/10.5771/2699-1284-2020-2-160).
- (2020b). “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: *Zeitschrift für digitale Geisteswissenschaften*. [10.17175/2020\\_006](https://doi.org/10.17175/2020_006).
- Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov (2017). “Membership inference attacks against machine learning models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. [10.1109/TDSC.2022.3180828](https://doi.org/10.1109/TDSC.2022.3180828).
- Song, Congzheng and Ananth Raghunathan (2020). “Information leakage in embedding models”. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. [10.1145/3372297.3417270](https://doi.org/10.1145/3372297.3417270).
- Song, Congzheng and Vitaly Shmatikov (2019). “Auditing Data Provenance in Text-Generation Models”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. Ed. by Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis. [10.1145/3292500.3330885](https://doi.org/10.1145/3292500.3330885).
- Thomas, Aleena, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow (2020). “Investigating the Impact of Pre-Trained Word Embeddings on Memorization in Neural Networks”. In: *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020*. Brno, Czech Republic: Springer, 273–281. [10.1007/978-3-030-58323-1\\_30](https://doi.org/10.1007/978-3-030-58323-1_30).
- Vulic, Ivan, Edoardo Maria Ponti, Robert Litschko, Goran Glavas, and Anna Korhonen (2020). “Probing Pretrained Language Models for Lexical Semantics”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*

- 2020, *Online*, November 16-20, 2020. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, 7222–7240. [10.18653/v1/2020.emnlp-main.586](#).
- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019a). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett. [10.5555/3454287.3454581](#).
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019b). “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. [10.18653/v1/W18-5446](#)".
- Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3. [10.1038/sdata.2016.18](#).



## A. Supplementary Material

Dataset	Precision				Precision				Precision							
	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent				
	<b>action<sub>100%</sub>: 5,903,010 lines</b>				<b>action<sub>10%</sub>: 590,818 lines</b>				<b>action<sub>1%</sub>: 59,330 lines</b>							
Action	0.9989	0.9986	0.9982	0.9931	0.9961	0.9954	0.9936	0.9769	0.9566	0.9479	0.9298	0.7899	0.6380	0.5794	0.4844	0.2182
Drama	0.9982	0.9979	0.9971	0.9889	0.9945	0.9934	0.9910	0.9668	0.9516	0.9418	0.9222	0.7702	0.6404	0.5820	0.4880	0.2211
Fluff	0.9982	0.9978	0.9970	0.9889	0.9943	0.9932	0.9908	0.9668	0.9541	0.9448	0.9262	0.7776	0.6419	0.5839	0.4899	0.2203
Gutenberg	0.9758	0.9709	0.9615	0.8693	0.9585	0.9502	0.9345	0.7842	0.8509	0.8229	0.7732	0.4430	0.4834	0.4141	0.3156	0.0945
	<b>drama<sub>100%</sub>: 4,854,969 lines</b>				<b>drama<sub>10%</sub>: 484,128 lines</b>				<b>drama<sub>1%</sub>: 48,569 lines</b>							
Action	0.9981	0.9977	0.9977	0.9884	0.9933	0.9920	0.9891	0.9601	0.9334	0.9201	0.8938	0.7041	0.5603	0.4939	0.3935	0.1641
Drama	0.9989	0.9986	0.9986	0.9931	0.9954	0.9944	0.9924	0.9723	0.9449	0.9339	0.9115	0.7597	0.5839	0.5193	0.4202	0.1825
Fluff	0.9981	0.9977	0.9977	0.9884	0.9936	0.9922	0.9895	0.9619	0.9407	0.9288	0.9054	0.7289	0.5812	0.5170	0.4178	0.1767
Gutenberg	0.9763	0.9716	0.9716	0.8725	0.9563	0.9476	0.9310	0.7710	0.8273	0.7956	0.7395	0.3986	0.4255	0.3535	0.2566	0.0739
	<b>fluff<sub>100%</sub>: 5,251,248 lines</b>				<b>fluff<sub>10%</sub>: 524,322 lines</b>				<b>fluff<sub>1%</sub>: 52,696 lines</b>							
Action	0.9972	0.9966	0.9954	0.9827	0.9912	0.9894	0.9856	0.9480	0.9280	0.9137	0.8853	0.6876	0.5841	0.5200	0.4211	0.1810
Drama	0.9973	0.9967	0.9956	0.9831	0.9914	0.9896	0.9859	0.9491	0.9328	0.9193	0.8928	0.7061	0.6025	0.5400	0.4425	0.1949
Fluff	0.9988	0.9986	0.9981	0.9928	0.9958	0.9949	0.9930	0.9746	0.9518	0.9421	0.9224	0.7756	0.6306	0.5713	0.4755	0.2172
Gutenberg	0.9726	0.9671	0.9564	0.8453	0.9495	0.9393	0.9201	0.7305	0.8126	0.7784	0.7186	0.3722	0.4415	0.3701	0.2724	0.0804
	<b>gutenberg<sub>100%</sub>: 2,728,188 lines</b>				<b>gutenberg<sub>10%</sub>: 273,263 lines</b>				<b>gutenberg<sub>1%</sub>: 27,240 lines</b>							
Action	0.9833	0.9798	0.9728	0.9007	0.9651	0.9579	0.9438	0.8104	0.8460	0.8163	0.7623	0.4446	0.3440	0.2685	0.1707	0.0496
Drama	0.9846	0.9814	0.9750	0.9089	0.9676	0.9608	0.9477	0.8223	0.8561	0.8280	0.7773	0.4686	0.3551	0.2793	0.1799	0.0525
Fluff	0.9806	0.9766	0.9687	0.8878	0.9616	0.9536	0.9383	0.7947	0.8451	0.8152	0.7618	0.4433	0.3510	0.2759	0.1765	0.0503
Gutenberg	0.9972	0.9966	0.9955	0.9744	0.9894	0.9873	0.9830	0.9110	0.8933	0.8727	0.8346	0.5041	0.3362	0.2627	0.1693	0.0421

Table 3: InvBERT Linear trained on every data sizes and evaluated across all eval datasets.

Dataset	Precision				Precision				Precision							
	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent	BLEU	3-gram	4-gram	Sent
	<b>action<sub>100%</sub>: 5,903,010 lines</b>				<b>action<sub>10%</sub>: 590,818 lines</b>				<b>action<sub>1%</sub>: 59,330 lines</b>				<b>action<sub>0.1%</sub>: 5,864 lines</b>			
Action	0.9797	0.9762	0.9692	0.9174	0.9443	0.9376	0.9190	0.8049	0.6792	0.6360	0.5522	0.4009	0.1253	0.0737	0.0338	0.0251
Drama	0.9675	0.9625	0.9521	0.8722	0.9290	0.9222	0.9000	0.7611	0.6679	0.6263	0.5418	0.3496	0.1274	0.0762	0.0352	0.0254
Fluff	0.9632	0.9579	0.9468	0.8625	0.9252	0.9186	0.8961	0.7547	0.6656	0.6241	0.5397	0.3468	0.1266	0.0761	0.0353	0.0256
Gutenberg	0.8299	0.8048	0.7664	0.5252	0.7284	0.6919	0.6358	0.3930	0.4009	0.3515	0.2631	0.1430	0.0608	0.0354	0.0114	0.0117
	<b>drama<sub>100%</sub>: 4,854,969 lines</b>				<b>drama<sub>10%</sub>: 484,128 lines</b>				<b>drama<sub>1%</sub>: 48,569 lines</b>				<b>drama<sub>0.1%</sub>: 4,796 lines</b>			
Action	0.9581	0.9508	0.9376	0.8356	0.9084	0.8934	0.8656	0.7058	0.5964	0.5513	0.4581	0.2789	0.1094	0.0628	0.0279	0.0195
Drama	0.9760	0.9719	0.9636	0.9050	0.9331	0.9219	0.8999	0.7813	0.6173	0.5689	0.4769	0.3093	0.1140	0.0658	0.0295	0.0206
Fluff	0.9589	0.9521	0.9396	0.8470	0.9122	0.8979	0.8715	0.7216	0.6033	0.5579	0.4654	0.2931	0.1129	0.0653	0.0295	0.0204
Gutenberg	0.8189	0.7928	0.7523	0.5084	0.6998	0.6610	0.6018	0.3696	0.3492	0.3036	0.2166	0.1206	0.0538	0.0266	0.0083	0.0096
	<b>fluff<sub>100%</sub>: 5,251,248 lines</b>				<b>fluff<sub>10%</sub>: 524,322 lines</b>				<b>fluff<sub>1%</sub>: 52,696 lines</b>				<b>fluff<sub>0.1%</sub>: 5,226 lines</b>			
Action	0.9501	0.9416	0.9262	0.8117	0.9050	0.8894	0.6104	0.6926	0.6079	0.5661	0.4757	0.2856	0.1161	0.0705	0.0325	0.0227
Drama	0.9551	0.9475	0.9333	0.8315	0.9112	0.8894	0.6104	0.7156	0.6185	0.5744	0.4851	0.3029	0.1204	0.0737	0.0343	0.0234
Fluff	0.9751	0.9709	0.9626	0.9064	0.9369	0.9265	0.6104	0.8007	0.6499	0.6013	0.5147	0.3416	0.1239	0.0749	0.0353	0.0243
Gutenberg	0.8015	0.7731	0.7289	0.4706	0.6502	0.6104	0.5499	0.3499	0.3497	0.3077	0.2215	0.1223	0.0541	0.0287	0.0091	0.0098
	<b>gutenberg<sub>100%</sub>: 2,728,188 lines</b>				<b>gutenberg<sub>10%</sub>: 273,263 lines</b>				<b>gutenberg<sub>1%</sub>: 27,240 lines</b>				<b>gutenberg<sub>0.1%</sub>: 2,755 lines</b>			
Action	0.9028	0.8904	0.8619	0.6513	0.8051	0.7848	0.7313	0.4522	0.3474	0.2736	0.1852	0.0915	0.0397	0.0161	0.0037	0.0049
Drama	0.9124	0.9019	0.8761	0.6813	0.8164	0.7976	0.7461	0.4766	0.3546	0.2809	0.1916	0.0975	0.0417	0.0172	0.0042	0.0050
Fluff	0.8964	0.8838	0.8545	0.6419	0.7993	0.7784	0.7246	0.4494	0.3450	0.2718	0.1836	0.0929	0.0396	0.0161	0.0038	0.0044
Gutenberg	0.9330	0.9241	0.9071	0.7528	0.8094	0.7839	0.7419	0.4917	0.2803	0.2175	0.1406	0.0764	0.0314	0.0141	0.0032	0.0070

Table 4: InvBERT Seq2Seq trained on every data size and evaluated across all eval datasets.