Evaluating Linformer's Performance on the WMT14 EN-DE Machine Translation Task

Marco Pampaloni
Department of Computer Science
m.pampaloni2@studenti.unipi.it
October 23, 2024

1. Introduction

The Transformer architecture, since its introduction in 2017 [Vas+17], has revolutionized the field of natural language processing (NLP), reaching the state of the art in various downstream tasks. Despite its massive parallelism capabilities, the Transformer struggles with longer sequence lengths due to its attention mechanism, which scales quadratically with the number of input tokens. This is a problem at both training and inference time.

For this reason, there has been a huge research effort in recent years to develop faster attention mechanisms, either exploiting the IO properties of the hardware these models run on [Dao+22], or by approximating the result of scaled dot product attention (SDPA).

The *Linformer* architecture [Wan+20] is an example of the latter approach. The authors first empirically show that the attention matrix is often low-rank, meaning that it could be approximated with an SVD decomposition by only keeping the largest singular values. This would of course introduce additional asymptotical complexity to the method, so the authors propose the adoption of linear projections on the keys and values matrices K, V to reduce their dimensionality and drive the computational complexity of the attention mechanism from $O(n^2)$ to $O(kn)$. The authors further show that the choice of k does not depend on the sequence length n, so that the scaling can be considered linear in n.

This work aims at replicating the results of [Vas+17] and [Wan+20] and comparing the two architectures on the WMT14 EN-DE machine translation task.

1.1. Multi Head Attention

The standard multi head attention (MHA) introduced by [Vas+17] is computed as follows

$$\begin{aligned}\text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1,...,\text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}\big(QW^Q, KW^K, VW^V\big) \\ &= \underbrace{\text{softmax}\left(\frac{QW^Q(KW^K)^T}{\sqrt{d_{\text{model}}}}\right)}_{P_i} VW^V\end{aligned}$$

1.2. Linformer Attention

The Linformer attention first projects the key and value matrices into a space of dimension $\mathbb{R}^{k\times d}$ with projection matrices $E, F \in \mathbb{R}^{k\times n}$ and then computes MHA as before:

$$\overline{\text{head}}_i = \text{Attention}\big(Q, E_i KW_i^K, F_i VW_i^V\big)$$

This produces an attention matrix $\overline{P}_i \in \mathbb{R}^{n\times k}$, which is computed in time linear with n. Since the projection matrices E, F are fixed in size before training, an obvious donwside of this approach is that the maximum sequence length n has to be known beforehand and the model cannot then scale to inputs with more tokens than this value. One workaround is to set the maximum sequence length n to a large number and handle

shorter inputs by slicing the projection matrices along their columns before multiplying them with the inputs K and V.

1.3. Masking

In a standard Transformer architecture, sequences are usually batched together in order to exploit the model's parallelism. This requires padding to be applied to the batch, but pad tokens should not contribute to the attention's output. This can be achieved by masking the attention matrix, zeroing out its elements in correspondence with input pad tokens. This if often done by setting each corresponding element to $-\infty$ prior to applying the row-wise softmax.

In Linformer this method cannot be applied, as the attention matrix is linearly projected to a lower-dimensional space. Instead, we apply masking by zeroing out elements in Q, K, V corresponding to pad tokens. This ensures that the pad tokens do not contribute to the final attention result.

1.4. Causal masking

The standard Transformer was trained on the WMT14 EN-DE machine translation dataset, adopting an encoder-decoder architecture. The self-attention mechanism in the decoder's layers needs to employ causal masking in order for future (right) tokens not to contribute to the output of past positions. In the Transformer this again can be achieved by masking, in this case by adding an upper triangular matrix filled with $-\infty$ values to the pre-softmax attention matrix.

Linformer, which was originally developed for encoder-only architectures, does not allow for causal masking, because however you mask the resulting pre-activation attention matrix, future tokens leak into the past due to the linear projections of the K and V matrices.

The Linformer attention mechanism thus cannot be applied in the self-attention layers of the decoder, while it can safely be used in the cross-attention stages because of the lack of causal requirements. This hinders the full scaling potential of the encoder-decoder Linformer architecture, which is empirically shown in Section 6.3 and Section 6.4.

2. Data

As in [Vas+17], we used the WMT14 EN-DE [Boj+14] dataset comprised of about 4.5 million sencente pairs. For ease of use, we used the data hosted on a Kaggle repository[1], which conveniently collects all the english to german sentences in a single CSV file for training. Additionally, it provides a validation (dev) and test dataset, although the validation set was discarded.

Because of resources constraints, we could not use or preprocess the entire dataset while keeping it in memory. Applying preprocessing on the fly during training would have slowed down the experiment significantly, rendering its results meaningless. Instead, we used half of the original training dataset and used the last 1% of it as a validation set.

We used the pretrained Huggingface[2] implementation of the BART [Lew+19] tokenizer, which employs Byte Pair Encoding (BPE) and was applied to the dataset prior to training as a preprocessing step. The vocabulary of the tokenizer included about 50k tokens, in contrast to the vocabulary size of 37k from [Vas+17]. We truncated each sentence in the dataset to 256 tokens, padding on the right when necessary to enable training in batches. This should not cause meaningful degradation in model performance since the average sequence length after tokenization of the training dataset is much lower for both source and target sentences. Table 1 shows post-tokenization statistics for the training dataset without padding applied.

[1] https://www.kaggle.com/datasets/mohamedlotfy50/wmt-2014-english-german

[2] https://www.huggingface.co

Input	Average Length	Min Length	Max Length
source	31	3	10864
target	59	3	8318

Table 1: Statistics for the training dataset, reporting average (rounded up to the nearest integer), minimum and maximum sequence lengths after tokenization for source and target sequences.

During training, each batch is randomly sampled and trimmed to the length of the longest non-padded sequence within the batch, significantly improving performance for batches with many short sequences (and, consequently, a high number of tokens).

The validation dataset is only used to log metrics during training and for scheduling the learning rate. The test dataset is instead used during the last evaluation step of each experiment.

3. Architecture

This work focused on two architectures: the vanilla Transformer model and the Linformer. Both employed an encoder-decoder structure following [Vas+17]:

- 6 stacked encoder layers composed by Multi-Head self-attention, followed by a pointwise multi layer perceptron;
- 6 stacked decoder layers composed by Masked Multi-Head self-attention, followed by Multi-Head cross-attention and a pointwise multi layer perceptron.

In the standard Transformer model, each attention layer employs the scaled dot-product attention (SDPA) introduced by [Vas+17]. Linformer instead uses the linearized attention mechanism proposed by [Wan+20] everywhere except in the self-attention stage of the decoder's layers, because of causality requirements (See Section 1.4).

The implementation of Linformer we provide employs shared projection matrices E, F across heads. This is one of the variants proposed by the authors [Wan+20].

Both the standard Transformer and the Linformer adopt residual connections and Layer Normalization applied after each attention block and feed forward layer as in [Vas+17]. Sinusoidal positional encoding and learned embeddings were used for both architectures.

Table 2 shows the hyperparameters that have been set for both models, including every variant of Linformer tested.

One last notable change made to the architectures is the choice of the vocabulary size: BART tokenizer's vocabulary has been resized to the next multiple of 8 in order to fully exploit the Tensor Cores[3] of the Nvidia GPU used during training (See Section 4), accelerating matrix products when their sizes are divisible by 8.

Parameter	Value	Description
d_{model}	512	Size of each embedding vector
h	8	Number of heads in multi-head attention (MHA)
d_k, d_v	64	Inner dimension of key and value vectors per head
d_{mlp}	2048	Hidden layer dimension of each pointwise MLP

Table 2: Hyperparameters shared by the tested models

[3]https://developer.nvidia.com/blog/programming-tensor-cores-cuda-9/

3.1. Text generation

In contrast to [Vas+17], autoregressive machine translation was implemented through greedy decoding [SVL14]. We limited the output's length to 256 tokens, but terminate early when each sequence in the batch has produced an EOS token.

3.2. Implementation

The Transformer and Linformer models were implemented using PyTorch `2.5.0a0+872d972e41` following the details presented in [Vas+17] and [Wan+20]. Inspired by PyTorch's and Hugginface's APIs, we adopted a compositional approach to model definition that allows to easily swap attention mechanisms between Transformer architectures created using the same general backbone structure.

On top of the standard Transformer, a series of useful wrappers have been implemented that allow the model to transparently tokenize input strings, embed tokens, apply positional encodings and produce output token distributions.

The following code defines an encoder-decoder Transformer architecture equivalent to that described in Section 3.

```
attn = MultiHeadAttention(dim=512, n_heads=8)
encoder = TransformerEncoder(TransformerEncoderLayer(dim=512, mlp_dim=2048, attn), n_layers=6)
decoder = TransformerDecoder(TransformerDecoderLayer(dim=512, mlp_dim=2048, attn), n_layers=6)
transformer = NLPTransformer(encoder = encoder, decoder = decoder, vocab_size = 50272)
transformer = LanguageModelingHead(transformer)
```

The `NLPTransformer()` module internally allocates learnable embeddings and applies positional encoding. The `LanguageModeleingHead()` module outputs a probability distribution over the vocabulary defined by the tokenizer and it provides basic generational functionalities such as greedy decoding. To define a Linformer model one can simply swap the `MultiHeadAttention()` module with `LinformerAttention()`. Note that although the defined APIs share similarities with those of PyTorch, they have been implemented from scratch, along with everything else except the BART tokenizer.

4. Hardware

The experiments were carried out locally on a Linux system running a single Nvidia RTX 3090 GPU with 24GB of GDDR6X VRAM and an Intel i7 4770k CPU overclocked at 4.4GHz. The system's memory amounted to 16GB of DDR3 RAM.

4.1. CPU bound

Given the dated system components, the experiments were bottlenecked by the CPU, which could not keep the GPU usage at 100% most of the time during training, hovering near the 96-98% range of utilization instead.

5. Experiments

Each experiment was conducted independently on the same machine within a Docker container based on Nvidia's PyTorch NGC[4] version 24.08, which provides an optimized set of libraries for efficient training and inference on GPUs. Additionally, mixed precision training[5] was employed throughout, reducing memory usage and significantly speeding up computations.

We trained each model using a negative log-likelihood loss function with teacher forcing, computed pointwise across the batch and averaged. At the end of each epoch, we logged the loss and perplexity score calculated

[4]https://catalog.ngc.nvidia.com/orgs/nvidia/containers/pytorch
[5]https://developer.nvidia.com/automatic-mixed-precision

Model	PPL (test)	BLEU (test)
Transformer	**3.41**	29.92
Linformer (k=32)	3.96	**30.08**
Linformer (k=64)	3.84	27.74

Table 3: Linformer performance against a vanilla Transformer model on the WMT14 EN-DE (test) dataset. The Linformer has slightly worse perplexity than the Transformer, but their BLEU scores are comparable.

on the validation dataset. For both the logging of training and validation metrics, as well as model checkpointing, custom callback routines were implemented using Weights & Biases[6]. These callbacks enabled automatic storage of metrics, system information and hyperparameters, and saved model weights whenever validation metrics improved during training.

Since the training sequence lengths varied due to padding trimming (see Section 2), we used a batch size of 88 input sequences. Based on the statistics reported in Table 1, this corresponds to an average of 2728 source tokens and 5192 target tokens per batch. In the worst-case scenario, the maximum number of source or target tokens per batch was 22528, which is comparable to the token count used in [Vas+17]. This configuration enabled high memory utilization without running into allocation errors.

The learning rate schedule followed a similar approach to [Vas+17], starting with a linear warmup over 4,000 steps from 10^{-7} to 10^{-4}. After the warmup, we applied a cosine annealing strategy, gradually decreasing the learning rate to 10^{-6} by the final epoch. Each experiment ran for 10 training epochs.

6. Results and Analysis

The following section presents the performance results of Linformer variants against a standard Transformer on the WMT14 EN-DE task.

6.1. Model performance

Table 3 shows that the Linformer performs comparably to the standard Transformer model on both tested metrics, scoring worse perplexities on the test dataset, but showing similar BLEU values. Even though perplexity intuitively drops as the parameter k grows, the BLEU score seems to worsen. This variation is somewhat expected since [Vas+17] actually performed their BLEU evaluations on the test dataset with an ensemble of models computed by averaging many training checkpoints, ultimately lowering variance across experiments.

6.2. Translation examples

In the following section, we provide a selection of translations generated from test samples, showcasing the performance of each model that was evaluated. This aims to offer a comparison of the translation quality produced by the different models tested. More translation examples can be found in Appendix A.

[6]https://wandb.ai/site/

Example Translation #1:

Linformer $k = 32$
Input: The school yard renovation was originally planned back in 2008/2009, however, high unplanned expenses meant that the work had to be pushed back.
Candidate: Die Schule-Renovierung wurde ursprünglich 2008/2009 geplant, wobei die hohen Kosten jedoch zu ungeplanten Kosten führen mussten.
Reference: Ursprünglich war die Schulhofsanierung sogar schon in den Jahren 2008/2009 geplant, doch hohe unplanmäßige Ausgaben brachten eine Verschiebung.

Linformer $k = 64$
Input: The school yard renovation was originally planned back in 2008/2009, however, high unplanned expenses meant that the work had to be pushed back.
Candidate: Die Schule wurde 2008/2009 geplant, aber eine hohe ungeplanten Ausgabenkosten, die die Arbeit zurückzudrängen musste, bedeutete eine ursprüngliche Renovierung.
Reference: Ursprünglich war die Schulhofsanierung sogar schon in den Jahren 2008/2009 geplant, doch hohe unplanmäßige Ausgaben brachten eine Verschiebung.

Transformer
Input: The school yard renovation was originally planned back in 2008/2009, however, high unplanned expenses meant that the work had to be pushed back.
Candidate: Die Renovierungsarbeiten waren ursprünglich im Jahr 2008/2009 geplant, hingegen mit hohen ungeplanten Ausgaben, die zurückzufahren mussten.
Reference: Ursprünglich war die Schulhofsanierung sogar schon in den Jahren 2008/2009 geplant, doch hohe unplanmäßige Ausgaben brachten eine Verschiebung.

Example Translation #2:

Linformer $k = 32$
Input: Consequently, they will be particularly motivated playing against their former coach.
Candidate: Sie werden deshalb auch gegen ihren ehemaligen Reisebus besonders motiviert sein.
Reference: Von daher werden sie gegen ihren Ex-Coach sicher ganz besonders motiviert sein.

Linformer $k = 64$
Input: Consequently, they will be particularly motivated playing against their former coach.
Candidate: Sie werden daher besonders motiviert gegen ihren früheren Busch spielen.
Reference: Von daher werden sie gegen ihren Ex-Coach sicher ganz besonders motiviert sein.

Transformer
Input: Consequently, they will be particularly motivated playing against their former coach.
Candidate: Deshalb werden sie besonders motiviert sein, gegen ihren ehemaligen Busch zu spielen.
Reference: Von daher werden sie gegen ihren Ex-Coach sicher ganz besonders motiviert sein.

6.3. Training time

The time required to fully train a model is of paramount importance when considering the cost of scaling models to billions of parameters and large context windows. Although fully applying a "linearized" attention mechanism throughout an encoder-decoder architecture is not possible, Linformer should still provide a measurable improvement in the wall-clock time required to train it compared to a standard Transformer. Table 4 shows the average time required to compute a training batch.

Linformer marginally improves training times, reducing them by about 7% when using $k = 32$ and by 5% when $k = 64$.

6.4. Inference time

To experimentally verify the computational efficiency of the Linformer architecture, we ran inference for each model variant on batches of randomly generated data, varying the sequence length while keeping the

Model	Time (s/batch)	Total Duration (hr)
Transformer	0.221	14.3
Linformer ($k = 32$)	**0.205**	**13.2**
Linformer ($k = 64$)	0.209	13.4

Table 4: Time required by the various model variants to compute a training step (forward and backward passes) and total experiment duration.

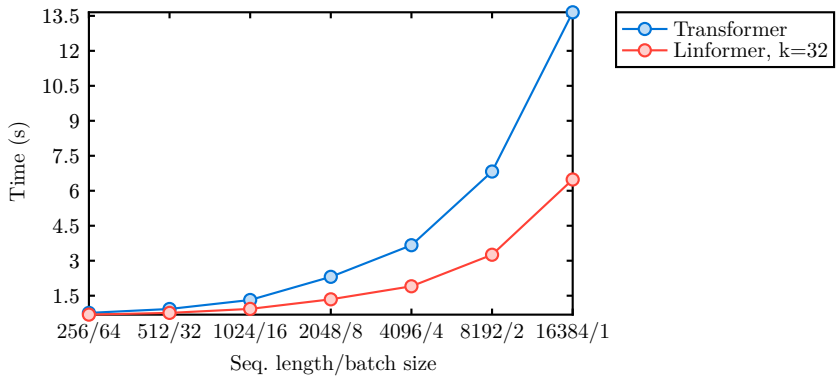

Figure 1: Comparison of scaling times between Linformer and Transformer with an encoder-decoder architecture. Different choices of the parameter k result in the same scaling, with differences in execution time falling within margin of error. This is because of the decoder's bottleneck which still requires the exact attention mechanism to be carried out.

total number of input tokens per batch constant. After a warmup phase, each inference step was repeated 10 times and the execution times were averaged. The same architectures described in Section 3 were used; however, in this case, we were not restricted to testing only trained models, as the output was discarded. The results of this analysis are summarized in Figure 1.

Despite Linformer's greater efficiency, the curves in Figure 1 indicate that the adopted encoder-decoder architecture does not scale linearly with the input sequence length, and in fact exhibits the same computational complexity as the Transformer model, up to a multiplicative constant. This is due to the use of the standard MHA in the decoder's self-attention layers.

In order to reveal the general scaling behaviour of the Linformer model, we carry out the same tests as before with an encoder-only architecture. The results are shown in Figure 2.

The wall-time plots clearly show that, as an encoder-only architecture, Linformer's computational complexity does not depend on the sequence length, which drastically improves inference performance as the context n grows, compared to a standard Transformer implementation.

7. Conclusions

Linformer's performance showed to be comparable with that of a standard Transformer on the WMT14 english to german machine translation task, while providing measurable efficiency gains over the former model, although they were limited by the decoder's bottleneck, which still required the use of the full attention matrix due to causality.

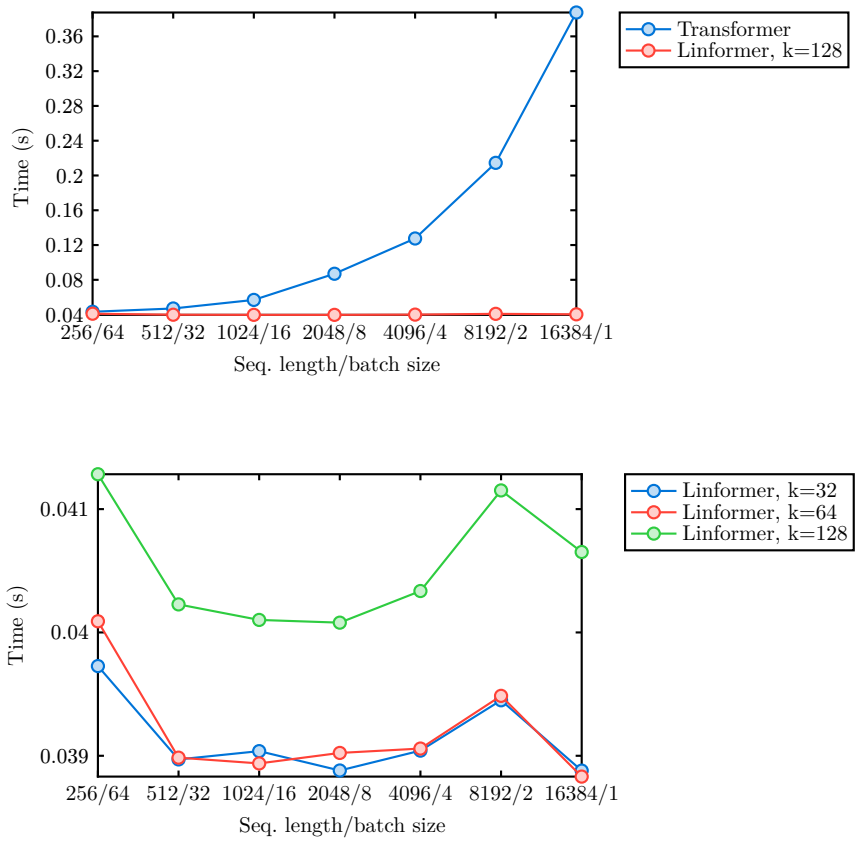

Figure 2: Scaling times of Linformer and Transformer with an encoder-only architecture. On the Top the standard Transformer is compared against the Linformer with $k = 128$. Linformer maintains constant execution time while varying the sequence. On the Bottom, various choices of parameters k are shown.

Investigating encoder-only architectures adopting Linformer's attention mechanism showed significantly faster inference speeds over the vanilla Transformer model, especially for longer sequences, suggesting that encoding tasks such as sentence classification might be a better benchmark for such models.

Encoder-decoder Transformers requiring causality masking could probably benefit more from other linearization approaches.

Bibliography

[Vas+17] A. Vaswani *et al.*, "Attention Is All You Need." [Online]. Available: http://arxiv.org/abs/1706.03762

[Dao+22] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness." [Online]. Available: http://arxiv.org/abs/2205.14135

[Wan+20] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-Attention with Linear Complexity." [Online]. Available: http://arxiv.org/abs/2006.04768

[Boj+14] O. Bojar *et al.*, "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, and L. Specia, Eds., Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 12–58. doi: 10.3115/v1/W14-3302.

[Lew+19] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." Accessed: Jul. 25, 2024. [Online]. Available: http://arxiv.org/abs/1910.13461

[SVL14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks." [Online]. Available: https://arxiv.org/abs/1409.3215

A. Translation Examples

Example #3

Linformer $k = 32$
Input: They're not all experienced racers, but people looking for excitement and adventure, and an achievable path towards world-class events.
Candidate: Sie sind nicht alle erfahrenen Rennen, sondern Menschen, die aufregend und ein erreichbares Erlebnis und ein erreichbares Weg zu den weltweit erreichten Veranstaltungen.
Reference: Sie sind nicht alle erfahrene Rennfahrer, sondern Leute, die auf der Suche nach Spannung und Abenteuer sind sowie nach einem erreichbaren Weg zu Weltklasse-Veranstaltungen.

Linformer $k = 64$
Input: They're not all experienced racers, but people looking for excitement and adventure, and an achievable path towards world-class events.
Candidate: Sie sind nicht alle erfahrenen Rennfahrer, sondern Menschen, die aufregende Erlebnisse und einen erreichbaren Weg zu Weltklasse-Events suchen.
Reference: Sie sind nicht alle erfahrene Rennfahrer, sondern Leute, die auf der Suche nach Spannung und Abenteuer sind sowie nach einem erreichbaren Weg zu Weltklasse-Veranstaltungen.

Transformer
Input: They're not all experienced racers, but people looking for excitement and adventure, and an achievable path towards world-class events.
Candidate: Sie sind nicht alle erfahrenen Rennen, sondern Menschen, die auf der Suche nach Spannung und Abenteuer und einem erreichbaren Weg in Richtung von Weltrangstouren suchen.
Reference: Sie sind nicht alle erfahrene Rennfahrer, sondern Leute, die auf der Suche nach Spannung und Abenteuer sind sowie nach einem erreichbaren Weg zu Weltklasse-Veranstaltungen.

Example #4

Linformer $k = 32$
Input: The letter extends an offer to cooperate with German authorities "when the difficulties of this humanitarian situation have been resolved."
Candidate: Das Schreiben erweitert ein Angebot an deutschen Behörden, "wenn die Schwierigkeiten dieser humanitären Situation gelöst wurden ".
Reference: In seinem Brief macht Snowden den deutschen Behörden ein Angebot der Zusammenarbeit, „wenn die Schwierigkeiten rund um die humanitäre Situation gelöst wurden“.

Linformer $k = 64$
Input: The letter extends an offer to cooperate with German authorities "when the difficulties of this humanitarian situation have been resolved."
Candidate: Das Schreiben erweist sich als Angebot an eine Zusammenarbeit mit deutschen Behörden "wenn die Schwierigkeiten dieser humanitären Situation gelöst wurden".
Reference: In seinem Brief macht Snowden den deutschen Behörden ein Angebot der Zusammenarbeit, „wenn die Schwierigkeiten rund um die humanitäre Situation gelöst wurden“.

Transformer
Input: The letter extends an offer to cooperate with German authorities "when the difficulties of this humanitarian situation have been resolved."
Candidate: Der Brief erweitert ein Angebot zur Zusammenarbeit mit deutschen Behörden ", wenn die Schwierigkeiten dieser humanitären Situation gelöst sind.
Reference: In seinem Brief macht Snowden den deutschen Behörden ein Angebot der Zusammenarbeit, „wenn die Schwierigkeiten rund um die humanitäre Situation gelöst wurden“.

Example #5

Linformer $k = 32$
Input: The residents of the Bischof-Freundorfer-Weg reported that a car was wrapped in toilet paper and its wheel trims stolen.
Candidate: Die Bewohner des Bischofs Freundorfer-Weg berichteten, dass ein Auto in der Papier- und ihrer Felgenkleidung gestohlen wurde.
Reference: Ein Anwohner im Bischof-Freundorfer-Weg meldete, dass sein Pkw mit Klopapier eingewickelt und seine Radzierblenden entwendet wurden.

Linformer $k = 64$
Input: The residents of the Bischof-Freundorfer-Weg reported that a car was wrapped in toilet paper and its wheel trims stolen.
Candidate: Die Bewohner des Bischof-Freundundund-Weg berichteten, dass ein Auto in der Toilette verpackt wurde und seine Felgen gestohlen wurde.
Reference: Ein Anwohner im Bischof-Freundorfer-Weg meldete, dass sein Pkw mit Klopapier eingewickelt und seine Radzierblenden entwendet wurden.

Transformer
Input: The residents of the Bischof-Freundorfer-Weg reported that a car was wrapped in toilet paper and its wheel trims stolen.
Candidate: Die Bewohner des Bischofs-Freundlöser-Weg berichteten, dass ein Auto in WC und das Rad für die Straßen gestohlen wurde.
Reference: Ein Anwohner im Bischof-Freundorfer-Weg meldete, dass sein Pkw mit Klopapier eingewickelt und seine Radzierblenden entwendet wurden.

Example #6

Linformer $k = 32$
Input: The ice cream harnesses the fluorescent properties of a jellyfish, synthesized by Chinese scientists
Candidate: Die Eiscen verwendet die fluessigen Eigenschaften eines Geleefish, synthetisiert durch chinesische WissenschaftlerInnen.
Reference: Die Eiscreme nutzt die fluoreszierenden Eigenschaften einer Qualle, die von chinesischen Wissenschaftler künstlich hergestellt werden

Linformer $k = 64$
Input: The ice cream harnesses the fluorescent properties of a jellyfish, synthesized by Chinese scientists
Candidate: Die Eiskreme nutzt die fluoreszen Eigenschaften eines Quallenhemfels, synthetisches, synthetisches, synthetisches, synthetisches.
Reference: Die Eiscreme nutzt die fluoreszierenden Eigenschaften einer Qualle, die von chinesischen Wissenschaftler künstlich hergestellt werden

Transformer
Input: The ice cream harnesses the fluorescent properties of a jellyfish, synthesized by Chinese scientists
Candidate: Die Eis-Creme kurbelt die fluoreszierenden Fluor-Eigenschaften eines Geleefis, die von den chinesischen Wissenschaftlern aus synthetisiert werden.
Reference: Die Eiscreme nutzt die fluoreszierenden Eigenschaften einer Qualle, die von chinesischen Wissenschaftler künstlich hergestellt werden

Example #7

Linformer $k = 32$
Input: Gutach: Increased safety for pedestrians
Candidate: Gutach: Stellverkehrssicherheit für Fußgänger
Reference: Gutach: Noch mehr Sicherheit für Fußgänger

Linformer $k = 64$
Input: Gutach: Increased safety for pedestrians
Candidate: Gutach: Sicherheit für Fußgänger
Reference: Gutach: Noch mehr Sicherheit für Fußgänger

Transformer
Input: Gutach: Increased safety for pedestrians
Candidate: Gutach: Wahrung der Sicherheit für Fußgänger
Reference: Gutach: Noch mehr Sicherheit für Fußgänger

Example #8

Linformer $k = 32$
Input: He was spared a driving ban on account of the special circumstances - but was not spared a fine.
Candidate: Er wurde ein Fahrverbot für die besonderen Umstände verschont, aber es wurde nicht gut erspart.
Reference: Ein Fahrverbot blieb ihm wegen der besonderen Situation erspart - die Geldstrafe aber nicht.

Linformer $k = 64$
Input: He was spared a driving ban on account of the special circumstances - but was not spared a fine.
Candidate: Er hat ein Fahrverbot aufgrund der besonderen Umstände verschont, aber keine schönen Worte geschont.
Reference: Ein Fahrverbot blieb ihm wegen der besonderen Situation erspart - die Geldstrafe aber nicht.

Transformer
Input: He was spared a driving ban on account of the special circumstances - but was not spared a fine.
Candidate: Er war aufgrund der besonderen Umstände ein Fahrverbot verschont, aber nicht ausgezeichnet.
Reference: Ein Fahrverbot blieb ihm wegen der besonderen Situation erspart - die Geldstrafe aber nicht.

Example #9

Linformer $k = 32$
Input: I did this mainly because I hankered after money and fame.
Candidate: Ich tat dies hauptsächlich, weil ich nach Geld und Ruhm hantschauen.
Reference: Ich habe es in erster Linie getan, weil ich auf Geld und Ruhm aus war.

Linformer $k = 64$
Input: I did this mainly because I hankered after money and fame.
Candidate: Ich tat dies hauptsächlich, weil ich nach dem Geld und Ruhm streiche.
Reference: Ich habe es in erster Linie getan, weil ich auf Geld und Ruhm aus war.

Transformer
Input: I did this mainly because I hankered after money and fame.
Candidate: Ich habe dies vor allem weil ich nach Geld und Ruhm.
Reference: Ich habe es in erster Linie getan, weil ich auf Geld und Ruhm aus war.

Example #10

Linformer $k = 32$
Input: This year, sexy inanimate objects are all the rage.
Candidate: Dieses Jahr, sexy in bewaffneten Objekten sind alle die Wut.
Reference: In diesem Jahr sind aufreizende unbelebte Gegenstände der letzte Schrei.

Linformer $k = 64$
Input: This year, sexy inanimate objects are all the rage.
Candidate: Dieses Jahr sind sexy in den dunklen Objekten die Wut.
Reference: In diesem Jahr sind aufreizende unbelebte Gegenstände der letzte Schrei.

Transformer
Input: This year, sexy inanimate objects are all the rage.
Candidate: In diesem Jahr sind sexy Objekte all die Wut.
Reference: In diesem Jahr sind aufreizende unbelebte Gegenstände der letzte Schrei.

Example #11

Linformer $k = 32$
Input: Was his death the result of a crime?
Candidate: War sein Tod das Ergebnis einer Verbrechen?
Reference: Ist sein Tod auf ein Verbrechen zurückzuführen?

Linformer $k = 64$
Input: Was his death the result of a crime?
Candidate: War sein Tod infolge eines Verbrechens?
Reference: Ist sein Tod auf ein Verbrechen zurückzuführen?

Transformer
Input: Was his death the result of a crime?
Candidate: War sein Tod das Ergebnis eines Verbrechens?
Reference: Ist sein Tod auf ein Verbrechen zurückzuführen?

Example #12

Linformer $k = 32$
Input: A "genuineness" test for foreign workers on 457 visas is being considered by the government as it contemplates expanding a crackdown.
Candidate: Ein "Gegensatz" für ausländische Arbeitnehmer auf 457 Visum wird von der Regierung als Regierung betrachtet, da sie ein hartes Vorgehen ausweitet.
Reference: Die Regierung denkt im Rahmen erweiterter Maßnahmen über einen „Echtheitstest“ für ausländische Arbeitnehmer mit 457-Visum nach.

Linformer $k = 64$
Input: A "genuineness" test for foreign workers on 457 visas is being considered by the government as it contemplates expanding a crackdown.
Candidate: Eine "genuinische Unzufriedenheit" für ausländische Arbeiter auf Visa wird von der Regierung als solche angesehen, da sie eine harte Aufhebung ausweitet.
Reference: Die Regierung denkt im Rahmen erweiterter Maßnahmen über einen „Echtheitstest“ für ausländische Arbeitnehmer mit 457-Visum nach.

Transformer
Input: A "genuineness" test for foreign workers on 457 visas is being considered by the government as it contemplates expanding a crackdown.
Candidate: Ein "Ganznachteil" Test für ausländische Arbeiter auf ein Visum für ausländische Arbeiter auf 457 wird von der Regierung als prägen, das sie eine härtere Hürde ausweitet.
Reference: Die Regierung denkt im Rahmen erweiterter Maßnahmen über einen „Echtheitstest“ für ausländische Arbeitnehmer mit 457-Visum nach.