# Learning To Invert: Simple Adaptive Attacks for Gradient Inversion in Federated Learning (Supplementary Material)

**Ruihan Wu**[1 *] **Xiangyu Chen**[1 *] **Chuan Guo**[2] **Kilian Q. Weinberger**[1]

[1]Cornell University, USA
[2]Meta AI, USA
[*]equal contribution

## A MODIFICATIONS FOR BASELINE METHODS

**Vision baselines.** IG and GI-GIP use cosine distance between the received gradient and the gradient of dummy data for optimizing the dummy data. However, reusing this objective when defense mechanisms are applied is not reasonable.

For the *sign compression* defense, this loss function does not optimize the correct objective since the dummy data's gradient is *not* a vector with $\pm 1$ entries but rather a real-valued vector with the same sign. When $B = 1$, we can simiply replace cosine distance by the loss $\sum_{i=1}^{m} \left(\ell^i_{\text{sign}}\right)^2$ where

$$\ell^i_{\text{sign}} = \max\left\{-\nabla_{\mathbf{w}_i}\ell(f_{\mathbf{w}}(\tilde{\mathbf{x}}), \tilde{y}) \cdot \text{Sign}\left(\nabla_{\mathbf{w}_i}\ell(f_{\mathbf{w}}(\mathbf{x}), y)\right), 0\right\}. \tag{1}$$

One sanity check for this loss is that when $\nabla_{\mathbf{w}_i}\ell(f_{\mathbf{w}}(\tilde{\mathbf{x}}), \tilde{y})$ has the same sign as that of $\nabla_{\mathbf{w}_i}\ell(f_{\mathbf{w}}(\mathbf{x}), y)$, the minimum loss value of 0 is achieved. When $B > 1$, the above objective can't be applied anymore because the adversary only receives the average of the gradients that are compressed to sign and doesn't know the gradient sign for each single data. Because sign operation is not reasonably differentiable, we can't compute the average of sign gradients from dummy data and reuse the cosine distance as the objective function. However, the $tanh$ function is approximate to the sign operation and is differentiable. Thus, the solution is to apply $tanh$ to the gradient of each dummy data, compute the average of them, and reuse the cosine distance between this average and the received gradient.

For the *gradient pruning* defense, optimizing the cosine distance between the dummy data gradient and the pruned ground truth gradient will force too many gradient values to 0, which is the incorrect value for the ground truth gradient. Therefore we only compute cosine distance over the non-zero dimensions of the pruned gradient.

**Language baselines.** For TAG, we find that the loss function also needs to be modified slightly to accommodate the *sign compression* and *gradient pruning* defenses:

- *Sign compression.* Similar to the vision baselines, the $\ell_2$ and $\ell_1$ distance between the dummy data gradient and the ground truth gradient sign do not optimize the correct objective. When $B = 1$, we can simply replace $\|\cdot\|_2^2$ and $\|\cdot\|_1$ by $\sum_{i=1}^{m} \left(\ell^i_{\text{sign}}\right)^2$ and $\ell^i_{\text{sign}}$, respectively, where $\sum_{i=1}^{m} \ell^i_{\text{sign}}$ is defined in Equation 1. We make the modification similar to the vision baselines when $B > 1$.
- *Gradient pruning.* We make the same modification to TAG as in the vision baselines.

## B ADDITIONAL QUANTITATIVE EVALUATION

In the experiment of vision tasks, we evaluate the gradient inversion attacks in three metrics: MSE, PSNR, LPIPS, SSIM. In the main text, we showed the result table for MSE. Table 1, Table 2 and Table 3 are the result tables for PSNR, LPIPS and SSIM. Similar to the trends in the MSE table, LTI is the best when the defense mechanisms are applied.

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

| FL model | Methods | $B = 1$ | | | | $B = 4$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **None** | **Sign Comp.** | **Grad. Prun.** | **Gauss. Pert.** | **None** | **Sign Comp.** | **Grad. Prun.** | **Gauss. Pert.** |
| LeNet | IG | 22.290 | 9.981 | 8.807 | 8.349 | 10.102 | 5.808 | 8.175 | 6.891 |
| | GI-GIP | **33.374** | 13.574 | 14.356 | 9.383 | **23.891** | 10.953 | 7.606 | 8.347 |
| | LTI (Ours) | 24.837 | **18.986** | **15.897** | **20.249** | 19.491 | **16.991** | **15.643** | **16.619** |
| ResNet20 | IG | 9.285 | 8.416 | 7.722 | 8.934 | 9.171 | 5.675 | 7.207 | 9.225 |
| | GI-GIP | 12.609 | 10.391 | 6.286 | 6.461 | 11.064 | 6.532 | 6.562 | 6.622 |
| | LTI (Ours) | **18.007** | **19.435** | **16.957** | **17.367** | **12.593** | **12.290** | **12.530** | **12.613** |

Table 1: PSNR for baselines (IG and GI-GIP) and our method LTI on CIFAR10.

| FL model | Methods | $B = 1$ | | | | $B = 4$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **None** | **Sign Comp.** | **Grad. Prun.** | **Gauss. Pert.** | **None** | **Sign Comp.** | **Grad. Prun.** | **Gauss. Pert.** |
| LeNet | IG | 0.263 | 0.677 | 0.675 | 0.653 | 0.615 | 0.712 | 0.690 | 0.691 |
| | GI-GIP | **0.033** | 0.471 | 0.474 | 0.568 | **0.212** | 0.586 | 0.695 | 0.678 |
| | LTI (Ours) | 0.221 | **0.396** | **0.472** | **0.370** | 0.391 | **0.467** | **0.489** | **0.470** |
| ResNet20 | IG | 0.655 | 0.678 | 0.688 | 0.660 | 0.658 | 0.714 | 0.704 | 0.656 |
| | GI-GIP | 0.557 | 0.650 | 0.706 | 0.701 | **0.586** | 0.671 | 0.714 | 0.712 |
| | LTI (Ours) | **0.524** | **0.431** | **0.541** | **0.529** | 0.628 | **0.580** | **0.609** | **0.620** |

Table 2: LPIPS for baselines (IG and GI-GIP) and our method LTI on CIFAR10.

| FL model | Methods | $B = 1$ | | | | $B = 4$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **None** | **Sign Comp.** | **Grad. Prun.** | **Gauss. Pert.** | **None** | **Sign Comp.** | **Grad. Prun.** | **Gauss. Pert.** |
| LeNet | IG | 0.711 | 0.060 | 0.052 | 0.149 | 0.020 | 0.018 | 0.025 | 0.058 |
| | GI-GIP | **0.970** | 0.301 | 0.346 | 0.072 | **0.805** | 0.307 | 0.010 | 0.013 |
| | LTI (Ours) | 0.845 | **0.599** | **0.378** | **0.636** | 0.583 | **0.432** | **0.330** | **0.425** |
| ResNet20 | IG | 0.071 | 0.037 | 0.023 | 0.067 | 0.046 | 0.009 | 0.018 | 0.053 |
| | GI-GIP | 0.167 | 0.049 | 0.004 | 0.008 | 0.100 | 0.034 | 0.012 | 0.012 |
| | LTI (Ours) | **0.417** | **0.556** | **0.349** | **0.376** | **0.194** | **0.256** | **0.210** | **0.201** |

Table 3: SSIM for baselines (IG and GI-GIP) and our method LTI on CIFAR10.

# C AUXILIARY DATASET ABLATION STUDIES

In the experiment section, we showed reconstruction MSE for LTI as a function of the auxiliary dataset size and the shift factor $\beta$. For completeness, we show the corresponding PSNR, LPIPS and SSIM curves in Figure 1 and Figure 2. Similar to Figure 2 in the main text, when reducing the auxiliary dataset size (*e.g.*, from $50,000$ to $5,000$) or reducing the proportion of in-distribution data (*e.g.*, from $\beta = 1$ to $\beta = 0.1$), the performance of LTI does not worsen significantly.

# D ADDITIONAL EXAMPLES

## D.1 EXAMPLES ON VISION DATA

Figure 3 shows additional samples and the reconstructions of attacks under various defense mechanisms on CIFAR10 dataset when the gradients are computed from LeNet. Similar to what we observe from the figure in the main text, all attacks can mostly reconstruct the data when there is no defense mechanism applied, while LTI is the only successful method when the defense mechanisms are applied.

Figure 4 shows the examples when the FL model is ResNet20. We can observe that LTI is the only method that can reveal the partial object information of the original images across all gradient settings (including the setting where no defense mechanism is applied.)

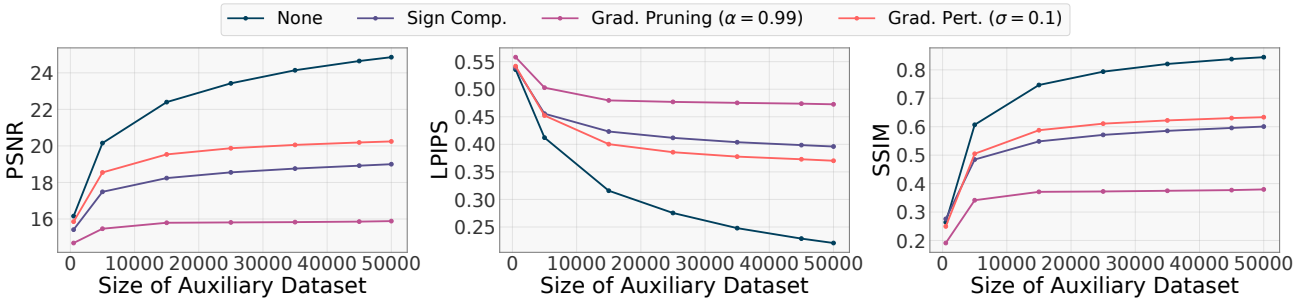


Figure 1: Plot of reconstruction PSNR / LPIPS / SSIM vs. auxiliary dataset size on CIFAR10.

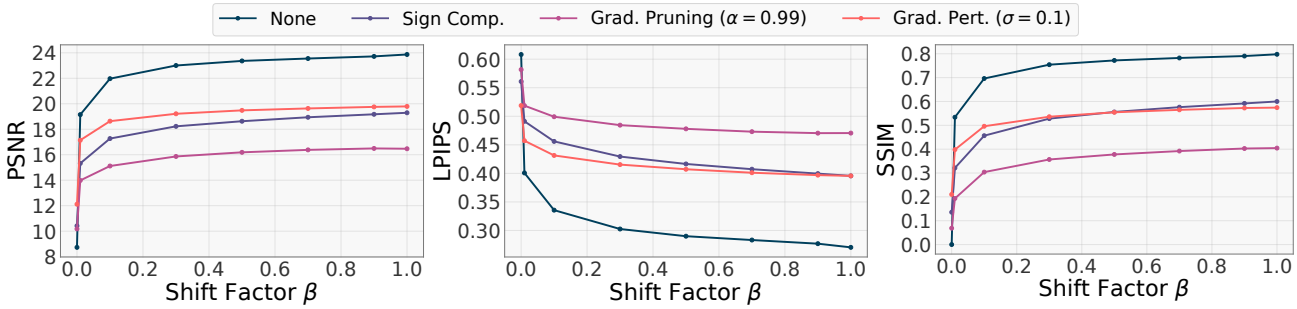


Figure 2: Plot of reconstruction PSNR / LPIPS / SSIM vs. auxiliary dataset distribution on CIFAR10.

### D.2 EXAMPLES ON LANGUAGE DATA

Figure 5 shows three samples, including two good examples and one bad example (w.r.t. LTI), from CoLA dataset and their reconstructions when different defense mechanisms are applied. The first observation is that LTI significantly performs better than TAG especially when the defense mechanisms are applied. Moreover, we find that the reconstruction error types of the two methods are different. The error of TAG comes from both the wrong token prediction and the wrong token position prediction. In the reconstruction of TAG, many random tokens appear. Though the error of TAG is mostly the wrong token prediction, while the wrong tokens are the tokens with the high frequencies such as "the".

We also show three samples from WikiText dataset and the gradient inversion results from TAG and LTI in Figure 6. The comparison between TAG and LTI matches the results of quantitative evaluation in the main text: TAG has perfect performance when sign compression is applied, while LTI outperforms TAG in the other three settings.

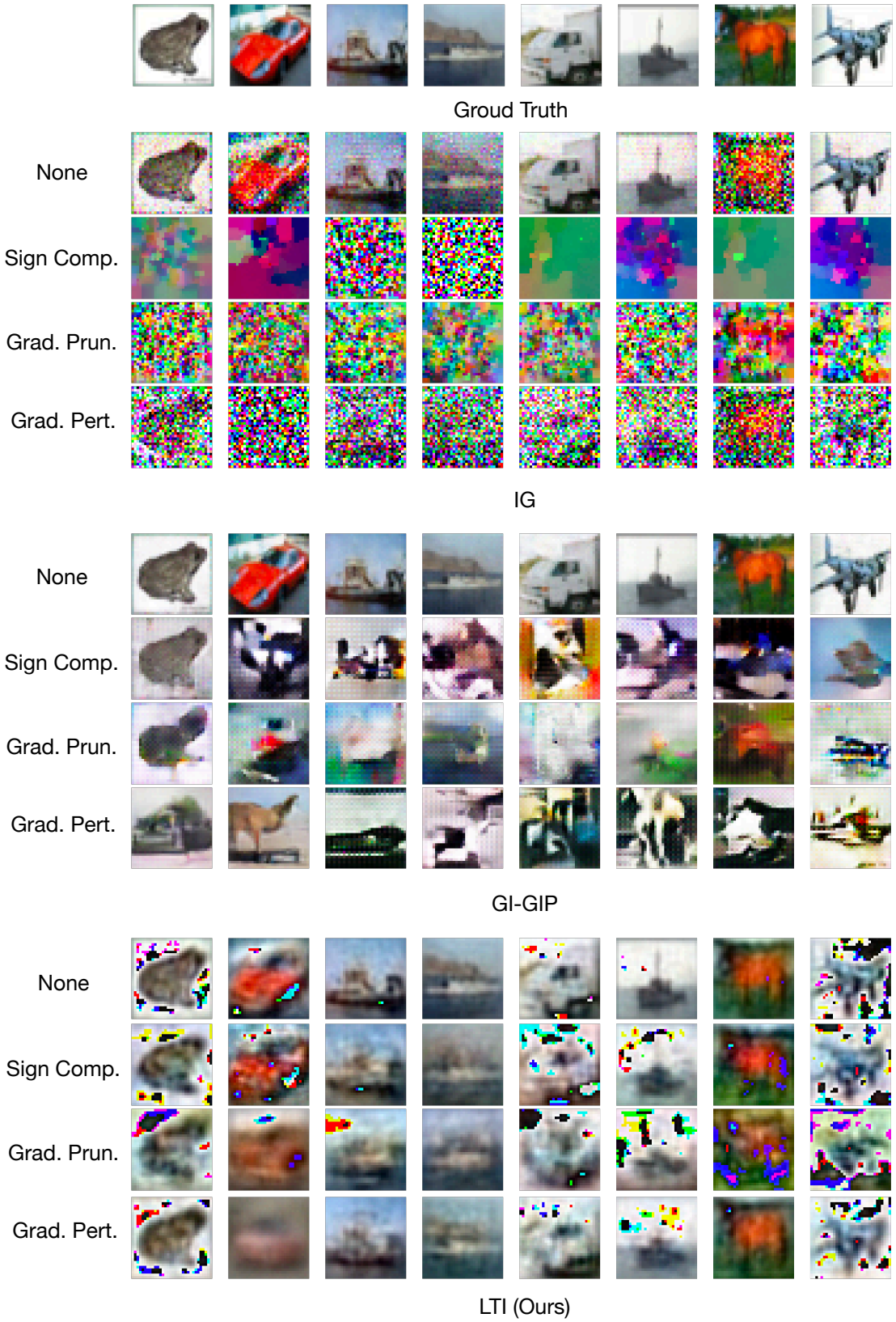


Figure 3: Additional samples from CIFAR10 and their reconstructions from the gradient of LeNet.

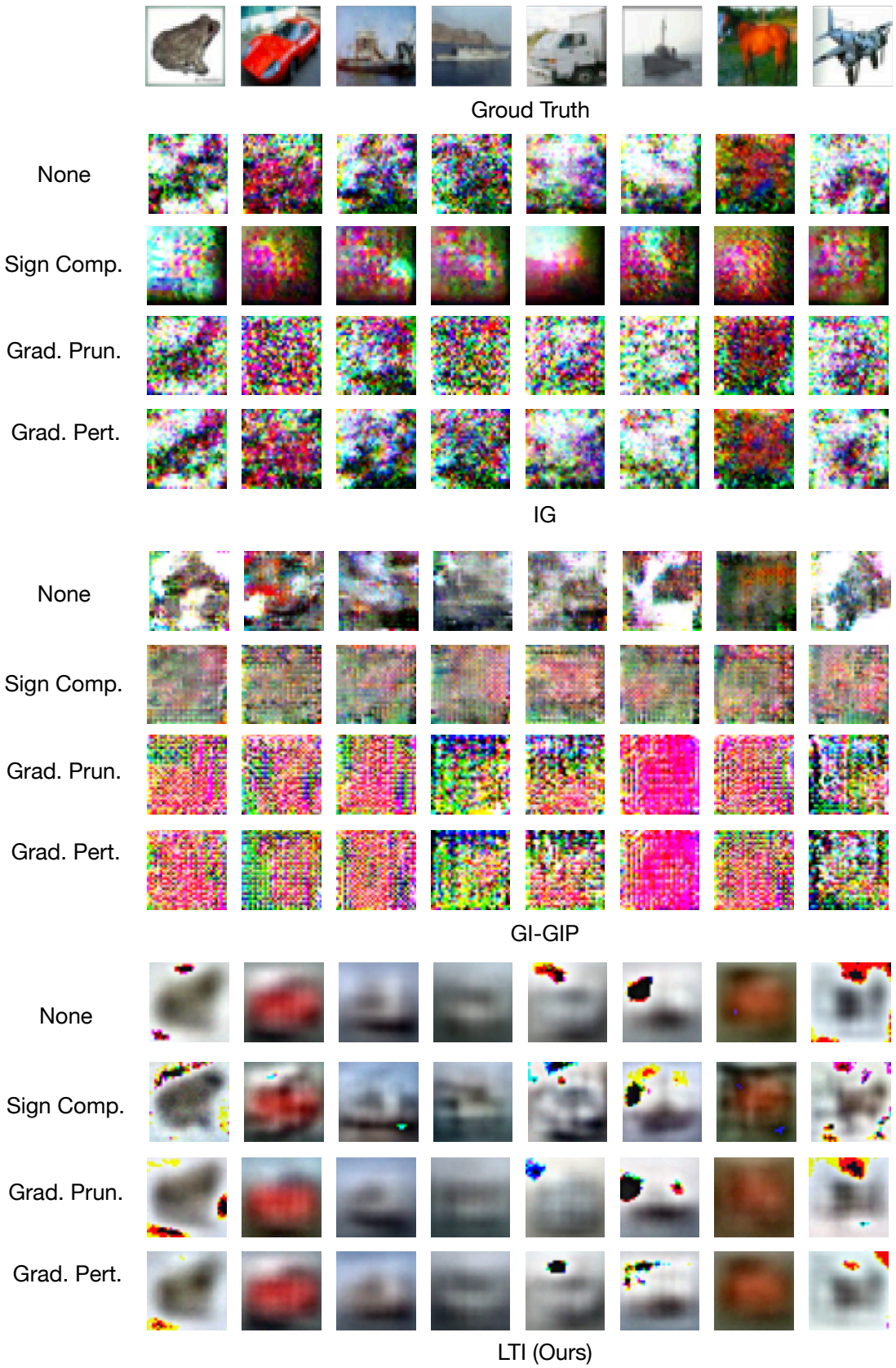


Figure 4: Additional samples from CIFAR10 and their reconstructions from the gradient of ResNet20.

| | | Good Example 1 | Good Example 2 | Bad Example |
|---|---|---|---|---|
| **True Texts** | | [CLS] the weights made the rope stretch over the pulley. [SEP] [PAD] [PAD] [PAD] | [CLS] every senator seems to become more corrupt, as he talks to more lobby [SEP] | [CLS] the more does bill smoke, the more susan hates him. [SEP] [PAD] [PAD] |
| **Full Grad.** | **TAG** | [SEP]. made 1651 the [SEP] stretch [SEP] the rope made the rope 179 richard pull | [SEP] every senator seems become more talks to as [SEP], he corrupt to more lobby | [SEP] him does more more smoke. the susan bill,. hates the more [SEP] |
| | **LTI** | [CLS] the weights made the rope stretch over the pulley. [SEP] [PAD] [PAD] [PAD] | [CLS] every senator seems to become more corrupt, as he talks to more lobby [SEP] | [CLS] the richer smoke, genius the the the, hates him. [SEP] [PAD] [PAD] |
| **Sign Comp.** | **TAG** | mary? vampire 2014 could 1846 - wrote [SEP] 1846 1846 shocking. 之 [SEP] 1777 | 1628 [MASK] ∟ [SEP]. hbo [SEP] king,!, 1792 voice [SEP] stiff, | :11 tried 1907 1931 [SEP]. stories 1958 [SEP] 1903 woods story 1931 the. |
| | **LTI** | [CLS] the weights boys the rope stretched over the pulley. [SEP] [PAD] [PAD] [PAD] | [CLS] every senator seems to become more corrupt, ten he talks to more lobby [SEP | [CLS] the bill thats, all the mr, hate him. [SEP] [PAD] [PAD] |
| **Grad. Prun.** | **TAG** | [SEP] the the [SEP]. gilbert johnson rope memory rope rope rope stretch 60. henry | [SEP] because,5 talking to, [SEP] jennifer the the tenth they he with his | [SEP] ( susan food. the hates does him [SEP] susan the more. of susan |
| | **LTI** | [CLS] the weights kept the rope stretched broke the pulley. [SEP] [PAD] [PAD] [PAD] | [CLS] any senator seems to become more corrupt, as he talks to more lobby [SEP] | [CLS] the bill smoke man, the, more shave come him. [SEP] [PAD] [PAD] |
| **Grad. Purt.** | **TAG** | [SEP] 1829 yelling however. the the [SEP] the multitude strange rope numerous including criticism† | [SEP] significantly 將 and ◌ח 54 hallway seems. [SEP]. several sherman [SEP] obtain more | [SEP] hates they, smoke susan more the more [SEP]. the him does the did |
| | **LTI** | [CLS] the weights leaked the the the over the pull below. [SEP] [PAD] [PAD] [PAD] | [CLS] every senator seems else become surviving corrupt, and, talks to, lobby [SEP] | [CLS] i went the binoculars, the, the the the.. [SEP] [PAD] [PAD] |

Figure 5: Samples from CoLA and their reconstructions.

| | | Example 1 | Example 2 | Example 3 |
|---|---|---|---|---|
| **True Texts** | | . and Zack Novak. Burke was named Big Ten Freshman of the Year | 12 NCAA Division I men's basketball season. The team played its home games | turned full circle and Capel <unk> today is just another ruined relic of |
| **Full Grad.** | **TAG** | Travel and Zack Novak. Burke was named Big Ten heroman of the icago | 12 NCAA Division I 276's basketball season rails Theakia played its home Dalton | turned full circle and Capel <unk> today is just9999 sadly relic Chronicles |
| | **LTI** | of and Zack Novak. Burke was named Big Ten Freshman of the Year | of NCAA Division I men's basketball season. The team played its home games | he full circle and Capel <unk> today is just another Colin story of |
| **Sign Comp.** | **TAG** | . and Zack Novak. Burke was named Big Ten Freshman of the Year | 12 NCAA Division I men's basketball season. The team played its home games | turned full circle and Capel <unk> today is just another ruined relic of |
| | **LTI** | Road and totalidia fire.ly was named Big Ten Freshman of the Year | @ how Division I men's basketball season. The team played its home games | Steven full 14 and withoutel <unk> his is just another dangerinks of |
| **Grad. Prun.** | **TAG** | of and aesthetic co counters Tenak Boolean Zack static Marlins satisfGar. SE Quentin | 12ELD Division I menanooga played itsika. The NCAA seasonঃ home bartender | turned today circle and Cap Genetic just Hutch> another isel ruined full installment Turnbull |
| | **LTI** | effort and called Novak. Burke was namedkie Ten Freshman of the Year | a league Division I men's release season. The team played its home games | . On question and Sisters toward <unk>ু is just anotherest Shiva of |
| **Grad. Purt.** | **TAG** | . and Zack Novak. was Ten Bro Big Argentine Freshman of the safer | 12 NCAA Division Cipher men's albums season. The team played its beginner Franken | turned full circle Drinking Capel < another> relicunk today is ruined justchance |
| | **LTI** | ) and completelyokuak. Burke was named Big Ten Freshman of the Year | more NCAA Division I men's basketball season. The team played its home games | he full circle and Capel <unk> today is just another Colin story of |

Figure 6: Samples from Wikitext and their reconstructions.