Perturbation Evaluation Set. For SQA and TABFACT, we also created new test sets to measure models' robustness to answer-invariant row and column perturbations during inference. Specifically, row and column orders are randomly perturbed for all tables in the standard test sets.[2]

Pre-training All the models are first tuned on the *Wikipidia* text-table pretraining dataset (Herzig et al., 2020), optionally tuned on synthetic dataset at an intermediate stage ("inter") (Eisenschlos et al., 2020), and finally fine-tuned on the target dataset. To get better performance on WTQ, we follow Herzig et al. (2020) to further pretrain on SQA dataset after the intermediate pretraining stage in the "inter-sqa" setting.

Evaluation For SQA, we report the cell selection accuracy for all questions (ALL) using the official evaluation script, cell selection accuracy for all sequences (SEQ), and the denotation accuracy for all questions (ALL_d). To evaluate the models' robustness in the instance level after perturbations, we also report a lower bound of example prediction variation percentage:

$$VP = \frac{(\text{t2f} + \text{f2t})}{(\text{t2t} + \text{t2f} + \text{f2t} + \text{f2f})} \quad (5)$$

where t2t, t2f, f2t, and f2f represents how many example predictions turning from correct to correct, from correct to incorrect, from incorrect to correct and from incorrect to incorrect, respectively, after perturbation. We report denotation accuracy on WTQ and binary classification accuracy on TABFACT respectively.

4.2 Baselines

We use TAPAS_BASE and $\text{TAPAS}_\text{LARGE}$ as baselines, where Transformer architectures are exactly same as BERT_BASE and BERT_LARGE (Devlin et al., 2019), and parameters are initialized from BERT_BASE and BERT_LARGE respectively. Correspondingly, we have our $\text{TABLEFORMER}_\text{BASE}$ and $\text{TABLEFORMER}_\text{LARGE}$, where attention bias scalars are initialized to zero, and all other parameters are initialized from BERT_BASE and BERT_LARGE.

4.3 Perturbing Tables as Augmented Data

Could we alleviate the spurious ordering biases by data augmentation alone, without making any modeling changes? To answer this, we train another set of models by augmenting the training data for TAPAS through random row and column order perturbations.[3]

For each table in the training set, we randomly shuffle all rows and columns (including corresponding column headers), creating a new table with the same content but different orders of rows and columns. Multiple perturbed versions of the same table were created by repeating this process $\{1, 2, 4, 8, 16\}$ times with different random seeds. For table QA tasks, selected cell positions are also adjusted as final answers according to the perturbed table. The perturbed table-text pairs are then used to augment the data used to train the model. During training, the model takes data created by one specific random seed in one epoch in a cyclic manner.

	Before Perturb			After Perturb	
	ALL	SEQ	ALL_d	ALL	VP
Herzig et al. (2020)	67.2	40.4	–	–	–
Eisenschlos et al. (2020)	71.0	44.8	–	–	–
Eisenschlos et al. (2021)	71.7	46.1	–	–	–
Liu et al. (2021)	–	–	74.5	–	–
TAPAS_BASE	61.1	31.3	–	57.4	14.0%
$\text{TABLEFORMER}_\text{BASE}$	**66.7**	**39.7**	–	**66.7**	**0.2%**
$\text{TAPAS}_\text{LARGE}$	66.8	39.9	–	60.5	15.1%
$\text{TABLEFORMER}_\text{LARGE}$	**70.3**	**44.8**	–	**70.3**	**0.1%**
TAPAS_BASE inter	67.5	38.8	–	61.0	14.3%
$\text{TABLEFORMER}_\text{BASE}$ inter	**69.4**	**43.5**	–	**69.3**	**0.1%**
$\text{TAPAS}_\text{LARGE}$ inter	70.6	43.9	–	66.1	10.8%
$\text{TABLEFORMER}_\text{LARGE}$ inter	**72.4**	**47.5**	**75.9**	**72.3**	**0.1%**

Table 1: Results on SQA test set before and after perturbation during inference (median of 5 runs). ALL is cell selection accuracy, SEQ is cell selection accuracy for all question sequences, ALL_d is denotation accuracy for all questions (reported to compare with Liu et al. (2021)). VP is model prediction variation percentage after perturbation. Missing values are those not reported in the original paper.

5 Experiments and Results

Besides standard testing results to compare TABLEFORMER and baselines, we also answer the following questions through experiments:

- How robust are existing (near) state-of-the-art table-text encoding models to semantic preserving perturbations in the input?
- How does TABLEFORMER compare with existing table-text encoding models when tested

[2] We fixed perturbation random seeds to make our results reproducible.

[3] By perturbation, we mean shuffling row and columns instead of changing/swapping content blindly.