Limitations of BERT:

- What NLP task cannot be solved w/ BERT?
  - text generation
  - translation, summarization, etc.

- Can we develop a self-supervised pretraining objective that covers all types of NLP tasks?
  - in addition to text gen, we'd like our model to handle classification, QA, sequence labeling, etc.

- TS paper: reformulate every NLP task as a text generation problem, "text-to-text"

input/output for TS:

students opened their books

masked LM:

students [MASK] their books

(predict "opened"

(Transformer Encoder)
The decoder is actually generating the text of the label, not a 3-way classification problem like in BERT.
TS Variants: (Table 2 of TS paper)

**Encoder-Decoder:**

- Input sequence: \( \langle x \rangle \) opened \( \langle y \rangle \)
- transformer encoder
- transformer decoder

- Output: \( \langle x \rangle \) teacher \( \langle y \rangle \) their book

**Decoder-only:**

- Next word prediction
- Transformer decoder
- Masked self-attention

- Input: \( \langle x \rangle \) opened \( \langle y \rangle \) [SEP] \( \langle x \rangle \) students ...

- Decoder LM also predicts next word for the input sequence
prefix LM!

Transformer decoder

next word prediction

partially-marked

self-attn

\[x\] opened \[y\] [SEP] \[x\] students ...

5) always
    unmarked
    never predicted