Different ways to use Transformer LMs:

- sequence-to-sequence models
- encoder/decoder models
- cross-attention
- encoder-only models
- BERT

Goal: translate French to English

Unmasked multi-head self-attention

Masked multi-head self-attention

<sos> e1 e2 e3

f1 f2 f3

Common Transformer Configurations:

1. Transformer encoder:
   - Transformer encoder:
     - [Diagram: Transformer encoder with labeled tokens C₁, C₂, C₃, C₄, and final-layer token-level representations.]

2. Transformer decoder:
   - Transformer decoder:
     - [Diagram: Transformer decoder with labeled tokens C₁, C₂, C₃, C₄, and predict symbols C₁, C₂, C₃, C₄.]

3. Transformer encoder/decoder:
   - [Text: Main differences:
     - Decoders can generate new text, encoders can’t.
     - Encoders observe a complete input; decoders only observe a prefix.
     - Encoders’ job is to produce powerful embeddings of the input; decoders’ is to generate.]
4. **Prefix LM**

```
<table>
<thead>
<tr>
<th></th>
<th>predict</th>
<th>predict</th>
<th>predict</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e_1</td>
<td>e_2</td>
<td>e_3</td>
</tr>
</tbody>
</table>
```

```
\uparrow
\quad\text{Partially-masked multi-head self-attention}
\quad\times L
\quad\text{Original decoder mask}
\quad\text{prefix LM mask}
```

```
f_1 \quad f_2 \quad f_3 \quad <\text{SOS}> \quad e_1 \quad e_2
```

---

**Pretraining an encoder-only Transformer:**

**LIBERT:** masked language modeling

```
<table>
<thead>
<tr>
<th></th>
<th>predict</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;Opened&quot;</td>
</tr>
</tbody>
</table>
```

```
\quad\text{Unmasked multi-head self-attention}
\quad\times L
```

```
[cls] \quad \text{the students} \quad \text{[MASK]} \quad \text{thru} \quad \text{books}
```

w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5
Instead of $p(w_3=\text{opened} | w_1, w_2)$
in MLM we have $p(w_3=\text{opened} | w_1, w_2, \text{[MASK]}, w_4, w_5)$

---

**how do we USE a masked LM?**

5) **Fine-tuning:** adjusting the params of a pretrained LM to adapt it to a single downstream task

**Ex: sentiment analysis:**

- Need a labeled dataset
- New softmax layer:

$$p = \text{softmax}(w_0 h_{\text{CLS}})$$

BERT

[CLS] this movie is good