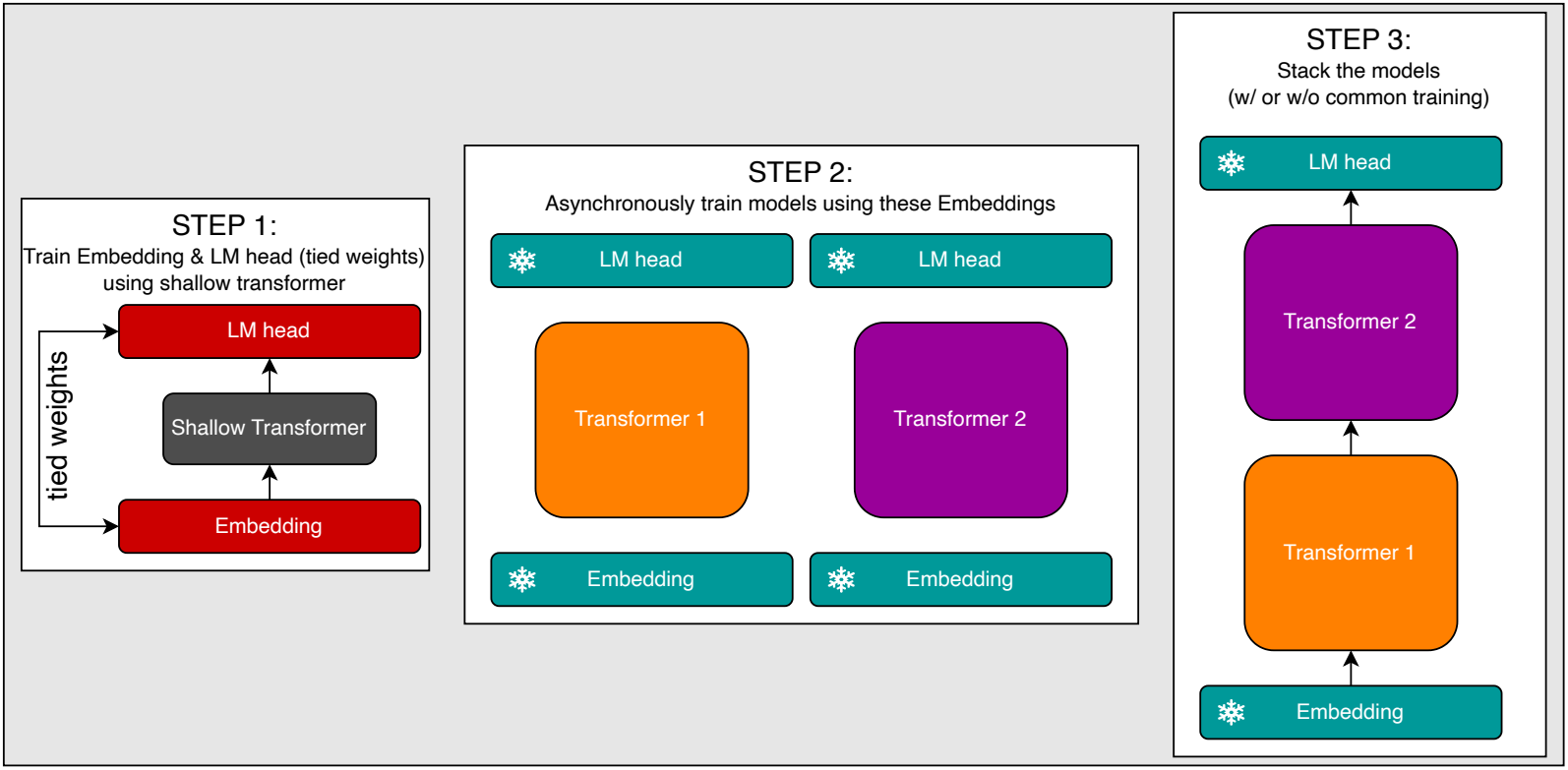
STEP 1:
Train Embedding & LM head (tied weights)
using shallow transformer
LM head
tied weights
Shallow Transformer
Embedding
STEP 2:
Asynchronously train models using these Embeddings
LM head
LM head
Transformer 1
Transformer 2
Embedding
Embedding
STEP 3:
Stack the models
(w/ or w/o common training)
LM head
Transformer 2
Transformer 1
Embedding