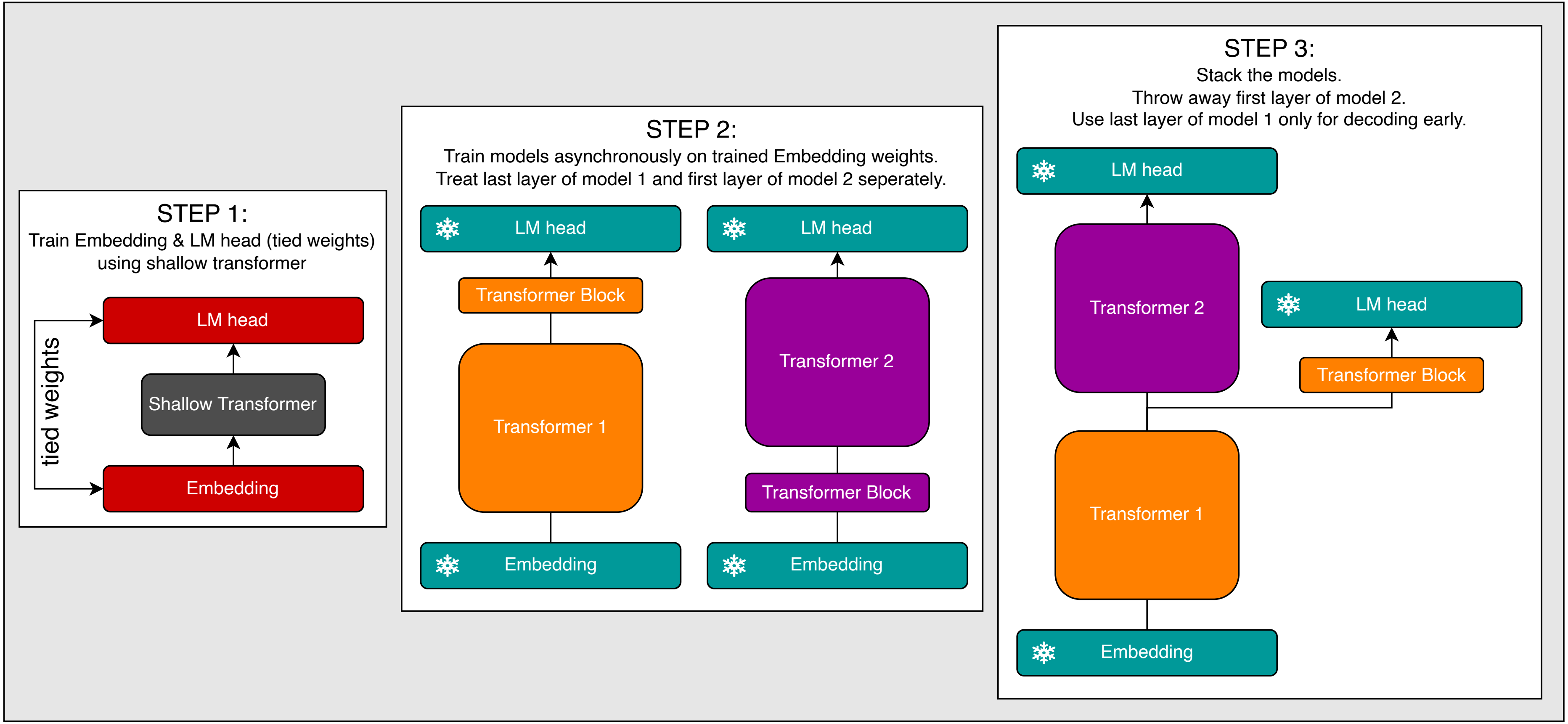

STEP 1:
Train Embedding & LM head (tied weights) using shallow transformer

STEP 2:
Train models asynchronously on trained Embedding weights.
Treat last layer of model 1 and first layer of model 2 seperately.

STEP 3:
Stack the models.
Throw away first layer of model 2.
Use last layer of model 1 only for decoding early.