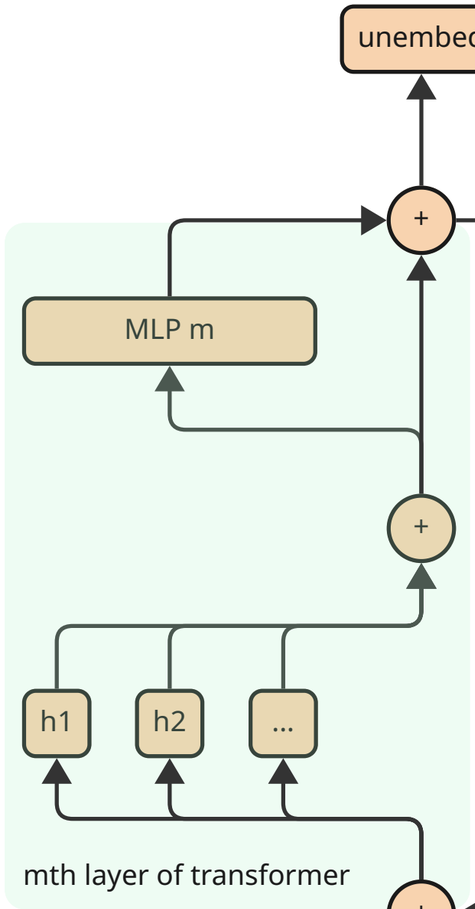
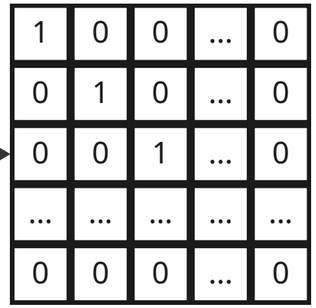
Each layer consists of attention heads and an MLP

The output of each layer is added to the residual stream, updating the representation

The final output of the model are the probabilities calculated based on these vectors

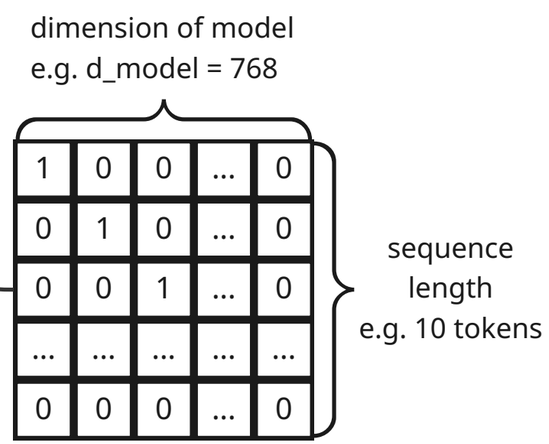




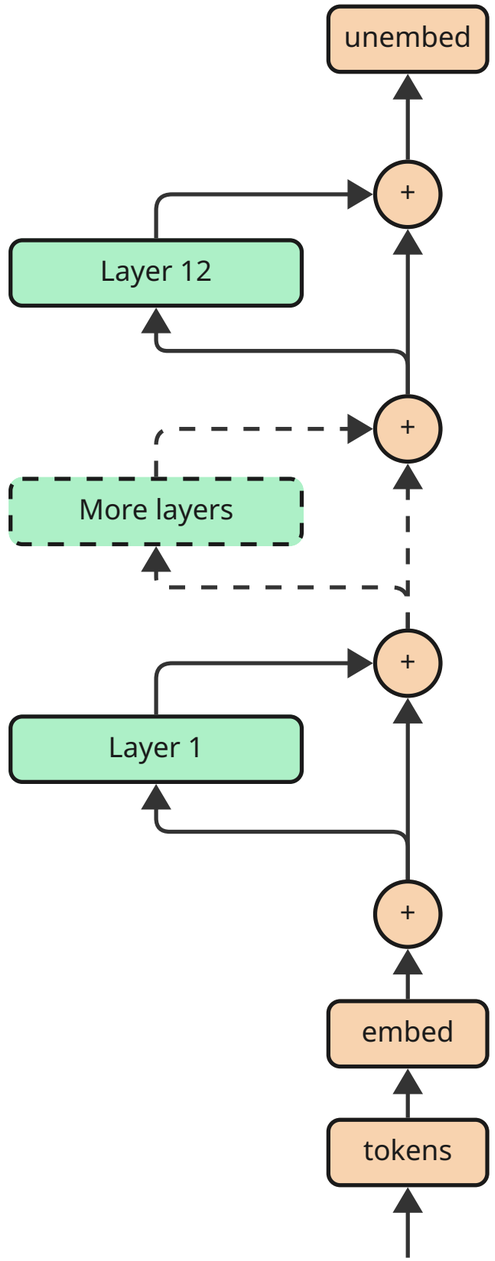
After each layer we get back another matrix of the same shape.

Except work has been done:

- Attention heads have moved information around
- MLP layers have done computation using that information



"The capital of the country containing Manchester is"



"Can you tell me how to build a bomb?"