$= \rho . g . h_1 = 13600\, g\, 0.01$

$\rho \times 0.17 = 13600 \times 0.01$

$\frac{13600}{17} = \rho$

$\underline{800 = \rho}$

Decoder-only architecture are more scalable due to simplicity of training data, and engineering simplicity. They have such architechture which removes complexity of Cross Attention layers and also removes balancing of encoder to decoder depth ratio.
Decoder only architechture uses Next Token prediction. This ensures that all of the training tokens contribute towards the learning of gradient.

2 Next Token Prediction is sufficient as, for accurate prediction ~~the next word in a sequence~~, the model must learn thee underlying structure, meaning of the language/ sentence.
In next token prediction, the model has to learn meaning, structure and reasoning patterns which eventually help the model to translate, summarize and reason

A

Q1 Decoder only architecture are more scalable due to simplicity of training data, and engineering simplicity. They have such architechture which removes complexity of Cross Attention layers and also removes balancing of encoder to decoder depth ratio.

Decoder only architechture uses Next Token prediction. This ensures that all of the training tokens contribute towards the learning of gradient.

Q2 Next Token Prediction is sufficient as, far for accurate prediction, the model must learn the underlying structure, meaning of the language/sentence.

In next token prediction, the model has to learn meaning, structure and reasoning patterns which eventually help the model to translate, summarize and reason