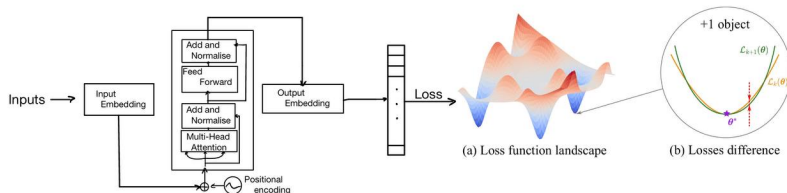


# Loss function Landscape Convergence for a Transformer Block



1. In the neighborhood of a local minimum, the loss function can be approximated by a quadratic form
2. When incrementally adding samples to the training set, we observe convergence in the optimization landscape

$$\mathcal{L}_k(\omega) \approx \mathcal{L}_k(\omega^*) + \frac{1}{2}(\omega - \omega^*)^\top \mathbf{H}^{(k)}(\omega^*)(\omega - \omega^*)$$

The described convergence will provide estimates on the minimum data size for efficient training.