MSE per sample on the test set using zero-shot model, fine-tuned and post discarding unreliable predictions

