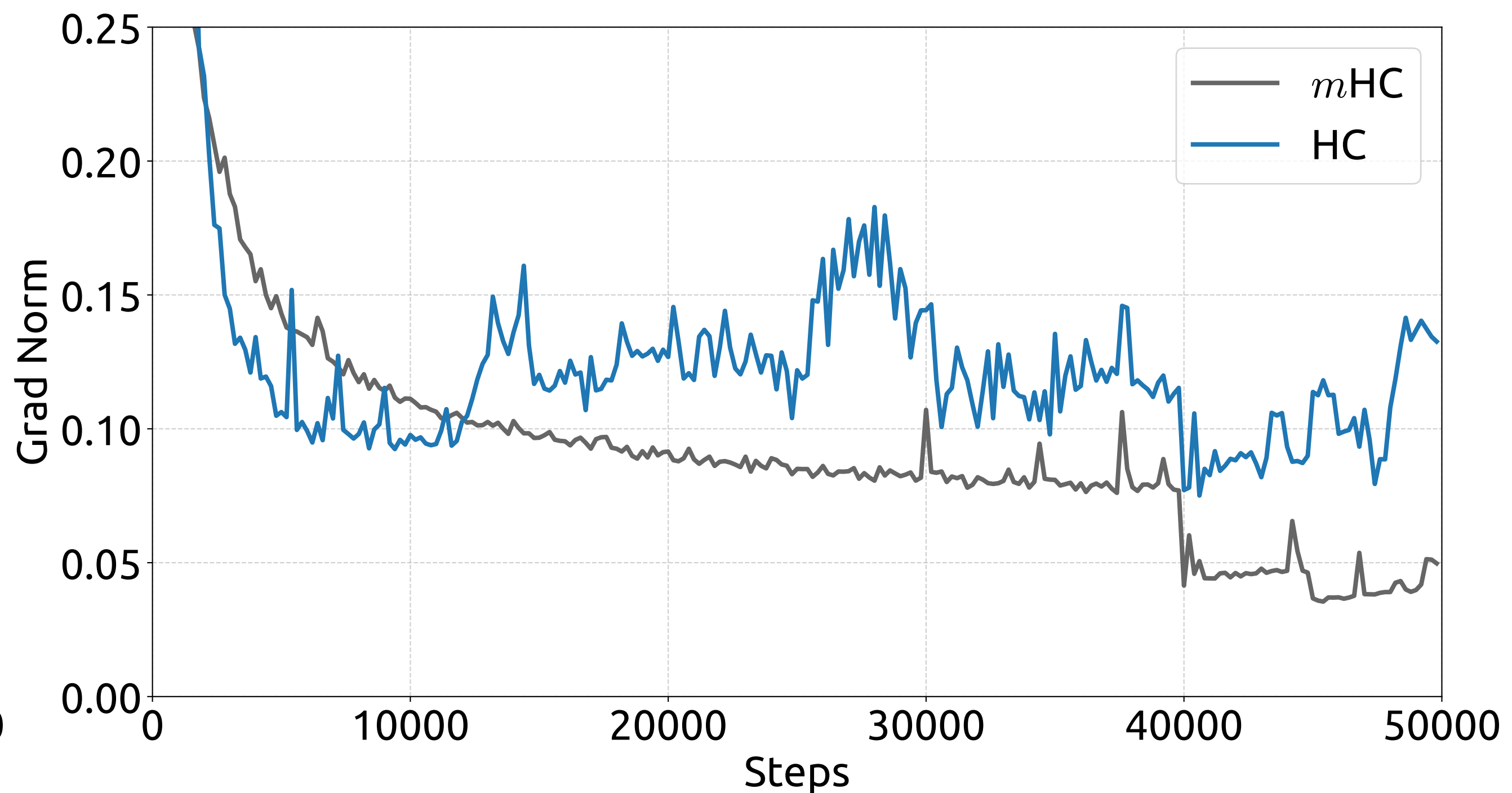(a) Absolute Training Loss Gap *vs.* Training Steps

(b) Gradient Norm *vs.* Training Steps