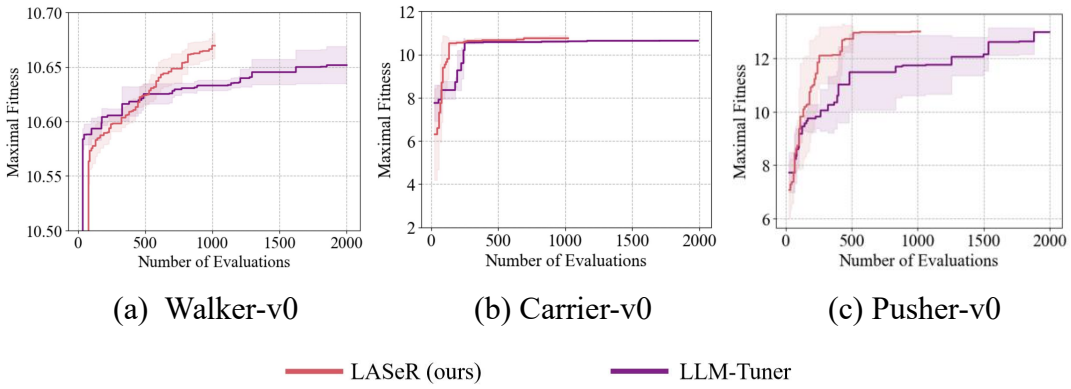
Supplementary material 1: Prolonged experiments of LLM-Tuner

We re-implemented LLM-Tuner, the most competitive baseline, for 2000 robot evaluations, which is double the number we employed in our original experiments. These experiments are intended to further verify that our rapid convergence is neither due to local optima or unreasonably easy task settings. Notably, as showcased in **Supplementary figure 1**, LLM-Tuner does not end up with higher fitness levels than those achieved by LASeR, largely confirming that our algorithm has not been stuck in local optima. Meanwhile, the evidently slower convergence of LLM-Tuner, which is especially pronounced in Walker-v0 and Pusher-v0), indicates that our fast convergence is more likely due to the effectiveness afforded by LLM-aided evolution and diversity reflection mechanism, rather than an artifact of task difficulty.



(a) Walker-v0 (b) Carrier-v0 (c) Pusher-v0

Supplementary figure 1. LASeR (red) compared with LLM-Tuner (purple), when the latter is implemented for 2000 robot evaluations. The colored bands represent (mean $\pm$ standard deviation). The results of LLM-Tuner are averaged across 3 independent runs, while those of LASeR are averaged over 5 runs.