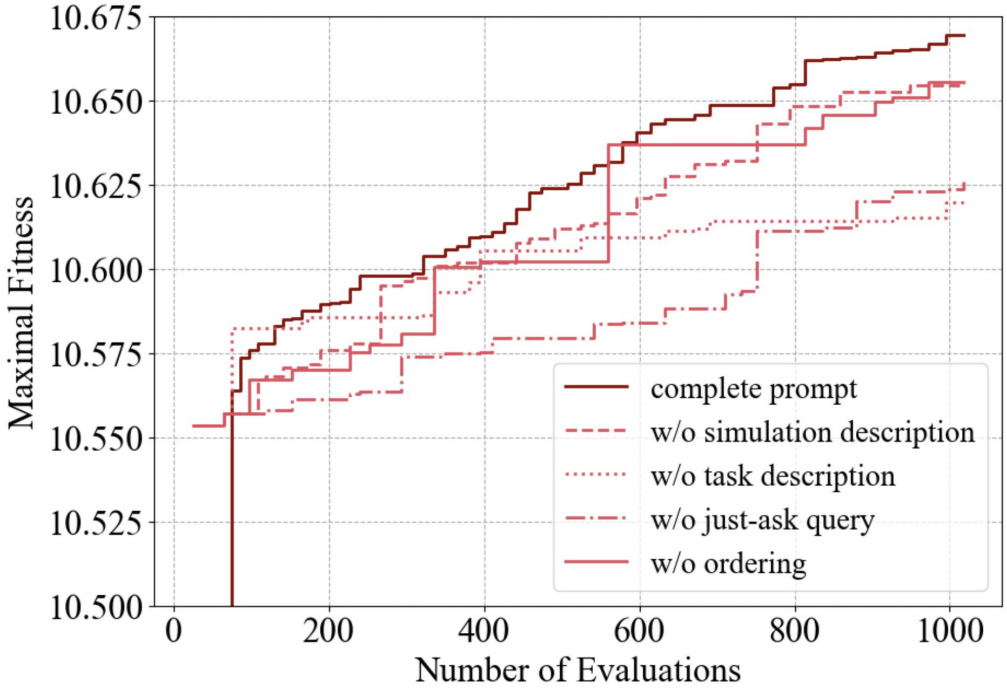
Supplementary material 10: Finer-grained ablation on prompt design

The prompt used in our study consists of three major components: task-related metadata, elite design-fitness pairs, and target fitness.

- The **task-related metadata** primarily includes descriptions of task objectives and the simulation environment. This component is largely derived from the official documents of EvoGym [], with minimal modifications. This metadata, which is often overlooked in previous works on LLM-aided robot design, serves two main purposes: to ground the evolutionary process in the specific context of the problem, and to facilitate the transfer of knowledge between different tasks.
- The second component consists of **elite design-fitness pairs** previously evaluated, where the designs are sorted according to their fitness in ascending order. This sorting is intended to leverage the pattern-completion capabilities of LLMs, a technique shown to be effective in prior research [].
- The third component, the **target fitness** (referred to as the **"just-ask query"** []), is introduced as a means of aligning the LLM's outputs with our desired results.

We have demonstrated the indispensability of task-related metadata in Section 3.3.2 of our paper. To further justify our prompt design and to complement the intuitive explanations provided above, we conducted finer-grained ablation studies and the results are reported in **Supplementary figure 10**. Specifically, we remove the following components one at a time: **(a)** the description of the simulation engine; **(b)** the description of task objectives; **(c)** the just-ask query (or target fitness); in this case, the LLM is simply prompted to generated robot designs with *higher* fitness; and **(d)** the ascending ordering of elite design-fitness pairs according to fitness. Our findings suggest that removing any of these components leads to performance drops. The just-ask query is proven the most essential, while simulation description and ordering play less important roles.

We would like to note that the phrasing of these components is intentionally left ***simple and intuitive***, without applying special techniques of prompt engineering. As such, our experimental results possess a certain level of robustness and do not hinge on the specifics of prompt designs. However, it would be a promising direction to integrate various prompting techniques, such as chain-of-thoughts and tree-of-thoughts, into our framework for better performances.



Supplementary figure 10. Finer-grained ablation studies on prompt design