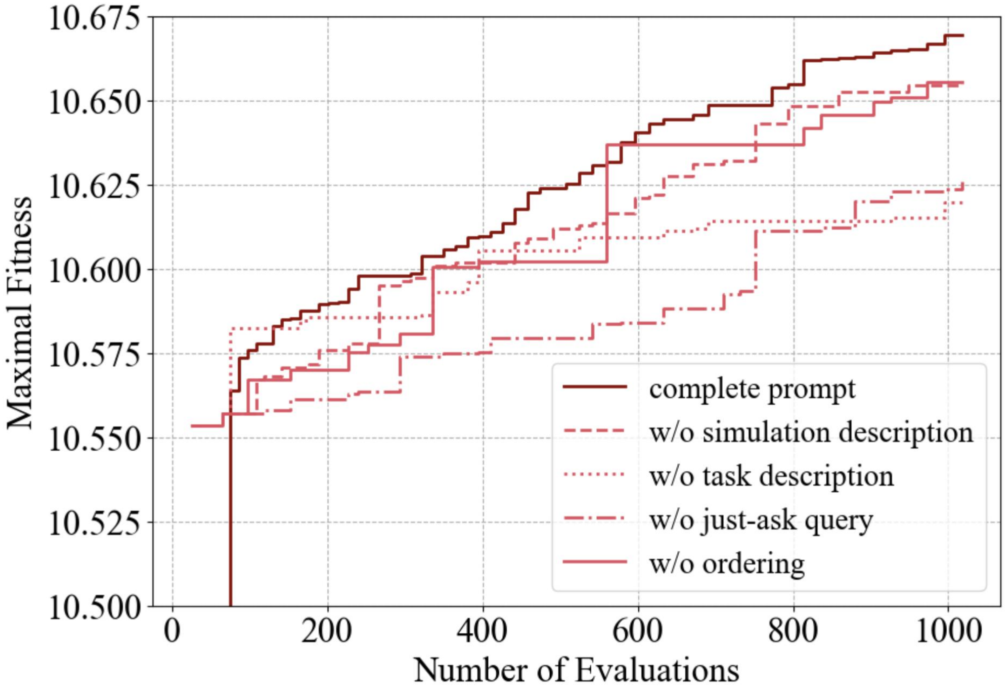
Supplementary material 10: Finer-grained ablation on prompt design

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 (Bhatia et al., 2021), 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 (Lange et al., 2024; Yang et al., 2024).
- The third component, the **target fitness** (referred to as the **"just-ask query"** in Lim et al. (2024)), 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-thought (Wei et al., 2022) and tree-of-thought (Yao et al., 2024), into our framework for better performances.



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

References:

[1] Bhatia, Jagdeep, et al. "Evolution gym: A large-scale benchmark for evolving soft robots." *Advances in Neural Information Processing Systems* 34 (2021): 2201-2214.

[2] Lange, Robert, Yingtao Tian, and Yujin Tang. "Large language models as evolution strategies." *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2024.

[3] Lim, Bryan, Manon Flageat, and Antoine Cully. "Large Language Models as In-context AI Generators for Quality-Diversity." *arXiv preprint arXiv:2404.15794* (2024).

[4] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

[5] Yang, Chengrun, et al. “Large Language Models as Optimizers.” *The Twelfth International Conference on Learning Representations*. 2024.

[6] Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." *Advances in Neural Information Processing Systems* 36 (2024).