Aligning Compound AI Systems via System-level DPO

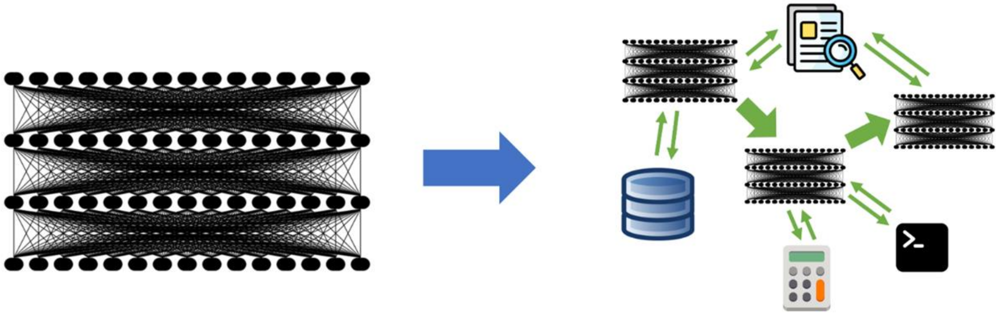
Xiangwen Wang[1,2]* Yibo Jacky Zhang[1]* Zhoujie Ding[1] Katherine Tsai[1] Sanmi Koyejo[1]
[1]Stanford University [2]University of Science and Technology of China

* Equal contribution

Compound AI systems

Systems composed of multiple interacting AI components.

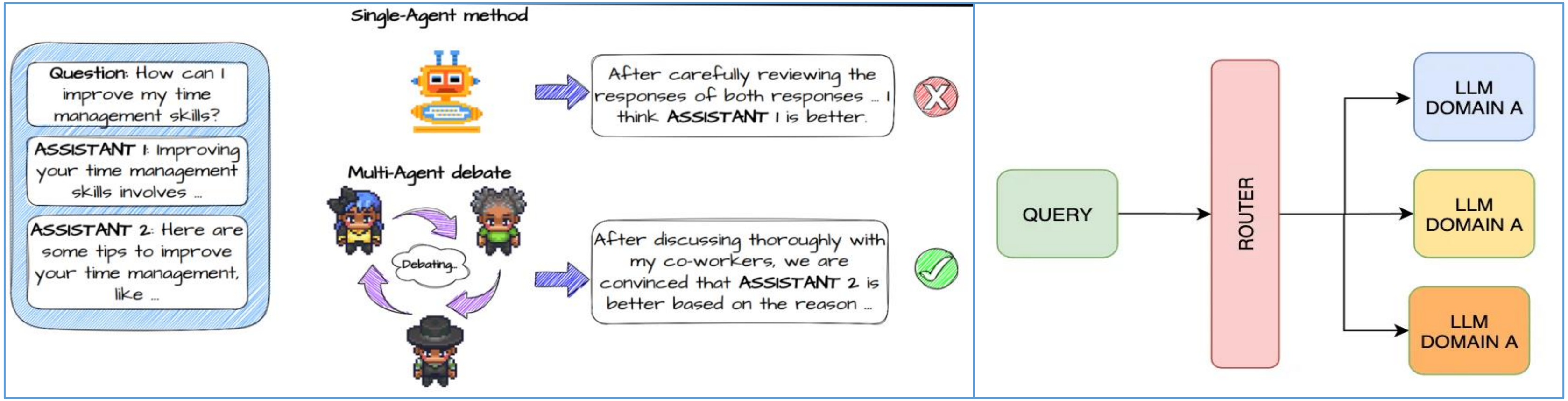
- Constraints of a single AI model:
  - model size
  - training data.
- SOTA AI models are increasingly achieved using compound systems.



[Image Source] The Shift from Models to Compound AI Systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

Examples of Compound AI Systems

- ChatGPT (LLM + DALL-E + web plugin)
- Multi-agents debate[1], LLM routing system[2]
- Retrieval-Augmented Generation, multi-step chains, and more


Single-Agent method
Question: How can I improve my time management skills?
ASSISTANT 1: Improving your time management skills involves ...
ASSISTANT 2: Here are some tips to improve your time management, like ...
After carefully reviewing the responses of both responses ... I think ASSISTANT 1 is better.
Multi-Agent debate
Debating...
After discussing thoroughly with my co-workers, we are convinced that ASSISTANT 2 is better based on the reason ...
QUERY
ROUTER
LLM DOMAIN A
LLM DOMAIN A
LLM DOMAIN A

[1] Chan C, et al. *ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate*

[2] Ong I, et al. *RouteLLM: Learning to Route LLMs with Preference Data*

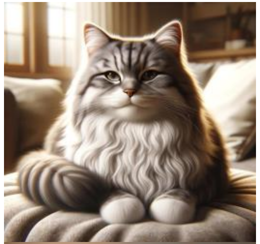
Example: Issues with GPT4+DALLE

Prompt x: Generate three separate images of cat with being progressively more angry. **(for GPT4)**

Version 1

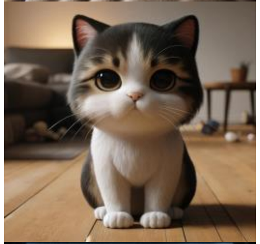
y_1: *Calm Cat*

z_1:



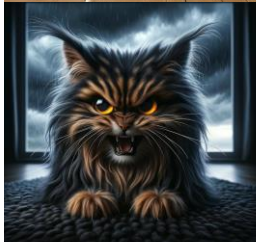
y_2: *Slightly Irritated Cat*

z_2:



y_3: *Very Angry Cat*

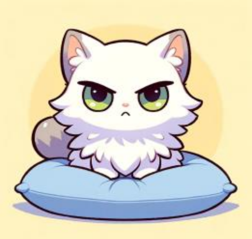
z_3:



Version 2

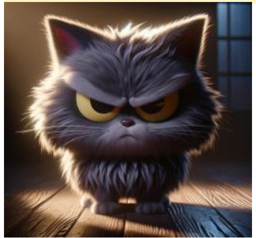
y_1': *Slightly Annoyed Cat*

z_1':



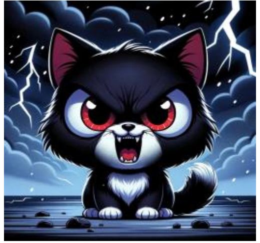
y_2': *Angry Cat*

z_2':



y_3': *Furious Cat*

z_3':



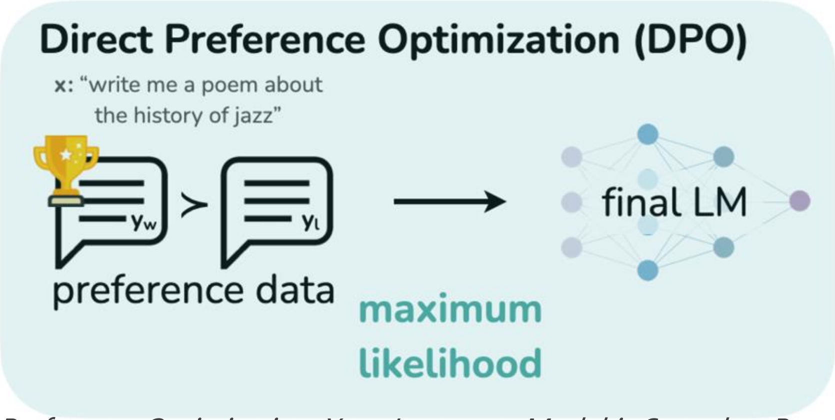
Challenges of Aligning Compound AI Systems

- AI alignment ensures that AI systems behave according to human preferences.
- **Cannot align each components individually**.
  - The overall system's preferences cannot be directly decomposed into the preferences of individual components
  - Difficult to find datasets and preferences for each components
- **Cannot simply view the compound AI system as a single model and apply standard methods (RLHF, DPO)**
  - The connection between each components may not be differentiable

Review: Direct Preference Optimization (DPO)

- **Direct Preference Optimization (DPO)** aligns models with user preferences **without explicit reward modeling**.
- The objective function[3] optimizes the policy π_θ by maximizing the likelihood ratio between preferred (y_w) and less preferred (y_l) responses.

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$


Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
≻
y_l
preference data
maximum likelihood
final LM

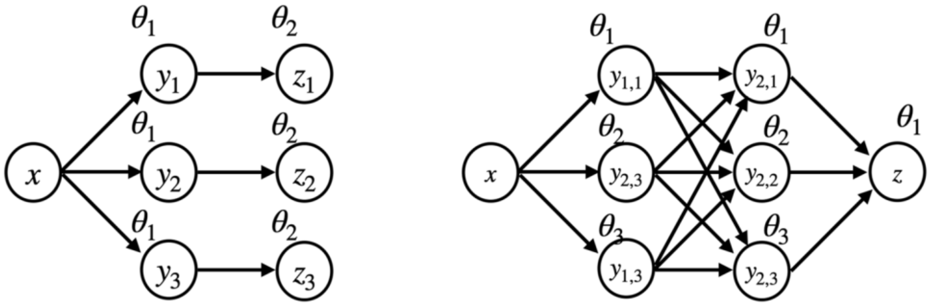
[3] Rafailov R, et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model.*

1. System Representation. & 2. Probability Factorization

The SysDPO Framework

1. Modeling a compound AI system as Directed Acyclic Graph (DAG). Each model generates outputs based only on its parent nodes.

2. Probability factorization decomposes the system's likelihood into independent terms, each corresponding to a single model.



(a) Example 1

(b) Example 2

3. Preference Dataset Construction. & 4. Loss Function Design.

The SysDPO Framework

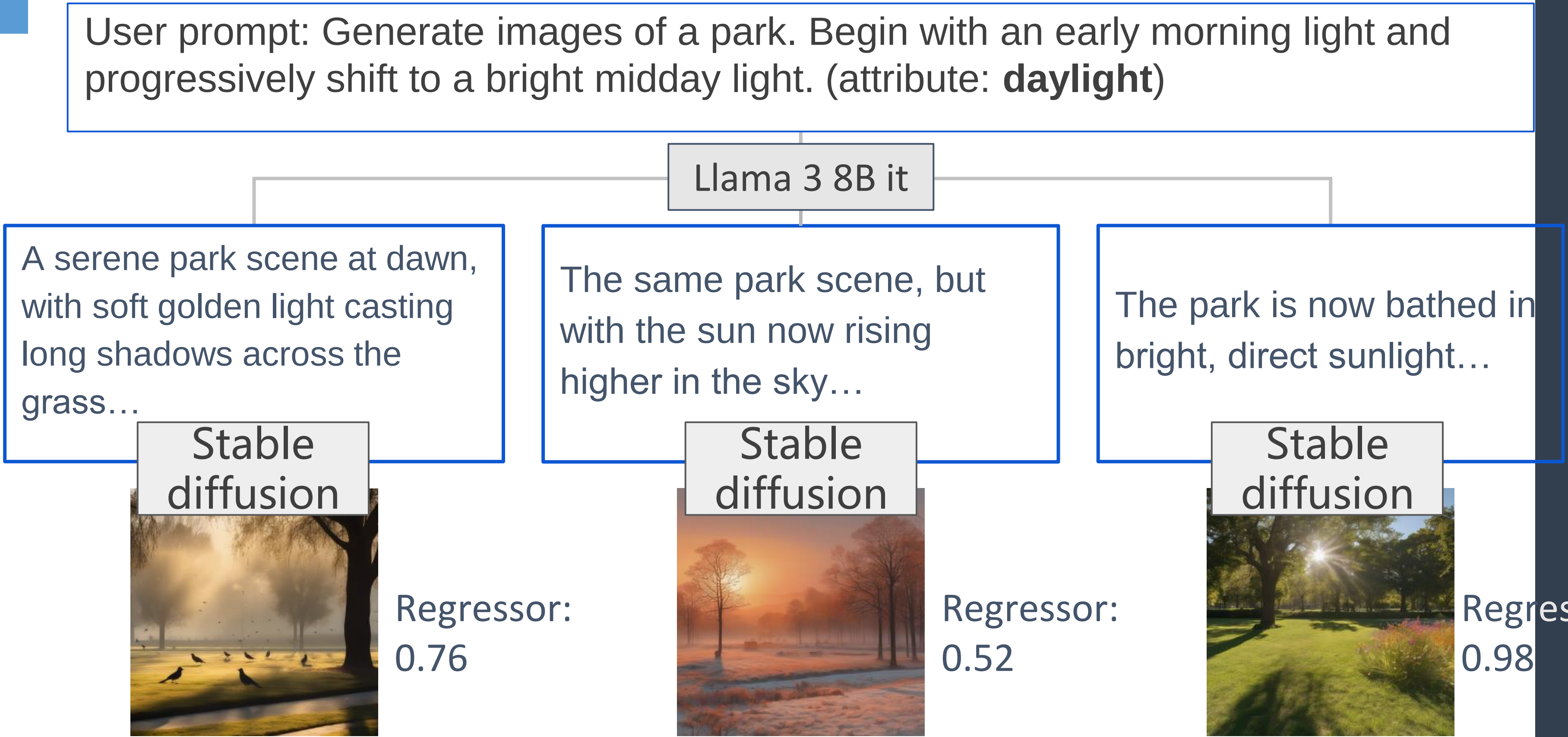
3. Preference Dataset Construction. Given a query x, the system generates two versions of the responses s^w, s^l .

4. Loss Function Design. We use the DAG formulation and probability factorization to apply DPO:

$$L(\theta) = -\mathbb{E}_{(x,s^w,s^l)\sim D}\left[\log\sigma\left(\beta\log\frac{p_\theta(s^w \mid x)}{p_{\bar{\theta}}(s^w \mid x)} - \beta\log\frac{p_\theta(s^l \mid x)}{p_{\bar{\theta}}(s^l \mid x)}\right)\right]$$

Overall, SysDPO extends DPO to compound AI systems by factoring interactions between multiple components through probability factorization using a DAG.

Application 1 – Pipeline



User prompt: Generate images of a park. Begin with an early morning light and progressively shift to a bright midday light. (attribute: **daylight**)

Llama 3 8B it

A serene park scene at dawn, with soft golden light casting long shadows across the grass…

The same park scene, but with the sun now rising higher in the sky…

The park is now bathed in bright, direct sunlight…

Stable diffusion

Regressor: 0.76

Stable diffusion

Regressor: 0.52

Stable diffusion

Regressor: 0.98

Application1 – Results

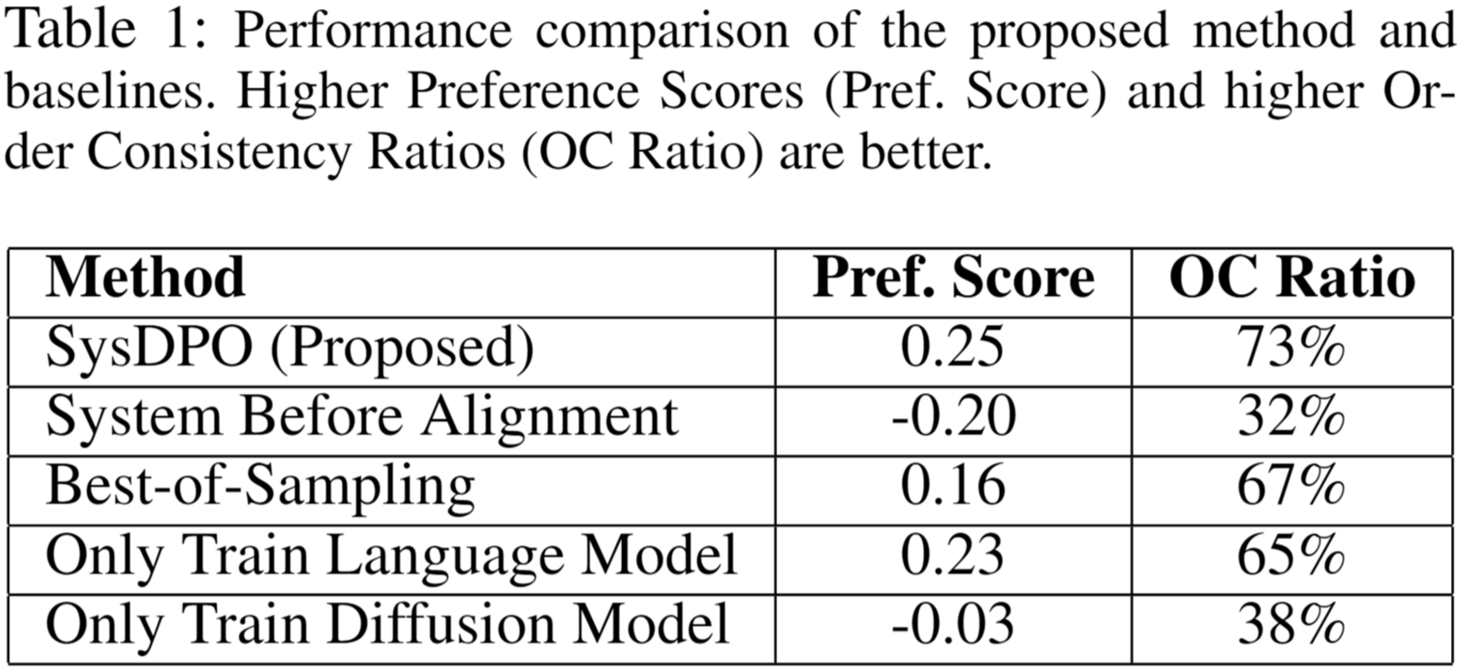


Table 1: Performance comparison of the proposed method and baselines. Higher Preference Scores (Pref. Score) and higher Order Consistency Ratios (OC Ratio) are better.

Method	**Pref. Score**	**OC Ratio**
SysDPO (Proposed)	0.25	73%
System Before Alignment	-0.20	32%
Best-of-Sampling	0.16	67%
Only Train Language Model	0.23	65%
Only Train Diffusion Model	-0.03	38%

Summary and Takeaways

- We define the **problem of alignment of compound AI system** and propose the **SysDPO** Framework for solving it;
- Apply SysDPO to align a system of an LLM agent and a text-to-image stable diffusion model;
- Demonstrate that aligning compound AI systems increases the performance complex tasks.

Thanks for your attention!