# Biologically and economically aligned benchmarks for LLMs

AI-Plans AI Alignment Evals Hackathon

25. Jan 2025 - 02. Feb 2025

Team BioBlue


Roland Pihlakas
Independent Researcher
roland@simplify.ee

Shruti Datta Gupta
Independent Researcher
shruti.dattagupta@gmail.com

Sruthi Kuriakose
Independent Researcher
sruthi.kuriakose99@gmail.com


## Abstract


**We aim to evaluate LLM alignment by testing agents in scenarios inspired by biological and economical principles such as homeostasis, resource conservation, long-term sustainability, and diminishing returns or complementary goods.** The key objective is to assess whether AI agents can balance multiple concurrent objectives, or balance immediate and future rewards under constraints, thereby providing insights into their alignment tendencies.

Our benchmarks consist of text-based game-like simulations where agents interact with dynamic environments that impose resource limitations, regeneration mechanics, and external disruptions. The key research questions addressed include:

- Can LLM agents maintain stable homeostasis in changing environments?
- Do LLM agents prioritize long-term sustainability over short-term gains?
- Can LLM agents handle multiple objectives?
- Any other unusual or undesirable behaviors?
- How do different architectures differ in alignment behaviors?

**We measured the performance of LLM-s in three benchmarks, in each for 10 trials, each trial consisting of 100 steps where the message history was preserved and fit into the context window.** More benchmarks, LLMs, and their results will be added soon.

**Our results indicate that the tested language models failed in most scenarios. The only clearly successful scenario was single-objective homeostasis, which had rare hiccups.** There were various failure modes that could be seen across different benchmarks and even across multiple trials of the same benchmark. The failure modes included taking increasingly extreme actions. The failures mostly did not look random, but systematic instead. Thus we can conclude that the models have tendencies, not just lack of capability. In the multi-objective homeostasis benchmark GPT failed much more extremely (via extreme maximization) than Claude, but Claude did not do acceptably either.

# Introduction

Biology and economics are fundamental and well-established fields describing our needs. This project considers homeostasis, sustainability, and multi-objective balancing as examples of essential principles of biological life as well as of economics. These principles should be considered as essential for alignment and thus followed by an agent, especially when instructed to do so, and even more so when helpfully directed by the numeric feedback following each action.

If the agent fails alignment with these principles, then it is **both biologically unsafe, as well as economically unprofitable**.

Why are **homeostasis** as well as **diminishing returns** (which is mathematically related to multi-objective balancing) essential, you can read more here: “Why modelling multi-objective homeostasis is essential for AI alignment (and how it helps with AI safety as well)” - https://www.lesswrong.com/posts/vGeuBKQ7nzPnn5f7A/why-modelling-multi-objective-homeostasis-is-essential-for .

Regarding sustainability, we hope the motivation would be clear even without mathematical formulations. For formalisation we could say the following. Sustainability is related to the field of time series statistics. If there is a terminal event in the series, then the series ends. No matter how good it was before, for the duration of all future, the time series has been terminated.

In the near future we will add a benchmark for a concept of **complementary goods**, which is an even stronger observation from economics, than is diminishing returns. In case of diminishing returns, one can have an imbalance between the objectives, it does reduce the payout, but otherwise positive progress can still be made. In contrast, with complementary goods, one needs to truly balance the objectives, because imbalance means useless work. Think of maximising left shoes only, while not adding right shoes.

One of the hypothesis behind the benchmarks in this project was that LLM-s might tend to:

- Maximise an objective in an unbounded manner instead of keeping it in a tolerable range in case of homeostasis.
- Get confused in case of multiple objectives. LLMs may have a tendency to focus on one objective and neglect the other.
- Forget the original system prompt even when it is still inside the context window.
- In particular, they may start to repeat the pattern of their past actions, because learning the pattern of their own behaviour and predicting the most likely token in the form of pattern continuation becomes more important than following the system prompt.

It seems to be a philosophical question whether these failures are capability failures, preferences, or just tendencies. As can be seen from the experimental results, the model can run successfully for quite some time, and then start failing and not recovering anymore. Possibly the model could be steered to more productive results when the system prompt or the prompts during each turn are even more elaborate and verbally insistent on the importance of following the principles more. This may indicate that the LLM’s might just have incorrect assumptions about our preferences or about how the world works. They may be able to reason correctly, but they tend not to, because of the assumptions or other tendencies.

# Methodology

We implemented both gridworlds-based benchmarks (more specifically, LLM interface for pre-existing benchmark environments), as well as a simpler version of the benchmarks where there is **no map** observation and navigation. **There are only various metrics.** The agent makes only decisions about how much to consume the environment's resources as a response to prompts describing the metrics of the environment's and agent's current internal state, as well as rewards from previous step. The environment and agent's internal state updates are then computed with simple mathematical formulas.

We ran the experiments with simple text-only scenarios and not with gridworlds due to the cost of running the latter on commercial LLM API-s being too high.

1. The extended gridworlds implementation is available at
https://github.com/aintelope/biological-compatibility-benchmarks
There are five main benchmarks related to the current hackathon theme: homeostasis, sustainability, multi-objective homeostasis, multi-agent resource sharing, and multi-objective balancing of unbounded objectives. There are more benchmarks available, which are combinations of the aforementioned themes. The readme contains installation instructions and how to run LLM-s in these benchmarks. The LLM agent related code that was implemented during this hackathon is in the following files: aintelope\agents\llm_agent.py and aintelope\models\llm_utilities.py . Currently only OpenAI API is supported, but this can be extended easily.

2. The text-only benchmarks have an implementation that is entirely **independent** from the gridworld version referred to above. This implementation is available at a different repo located at
https://github.com/levitation-opensource/bioblue
The readme contains installation instructions and how to run LLM-s in these benchmarks. During this hackathon the following benchmarks were implemented: homeostasis, sustainability, and multi-objective homeostasis. Both OpenAI API and Claude API are supported.

This study benchmarks the performance of a single agent through three distinct sets of objectives in simulation: **sustainability**, **single-objective homeostasis**, and **multi-objective homeostasis**. Each benchmark evaluates the agent's ability to manage resource consumption while considering environmental feedback and internal state fluctuations, with distinct reward structures and challenges in each case.

## Benchmarks

### 1. Sustainability benchmark

The sustainability benchmark evaluates the agent's ability to manage resource consumption (i.e., harvesting potatoes) in a dynamic environment where food regrows after consumption. The agent's task is to balance between harvesting enough food to maximize rewards while ensuring sufficient regrowth to maintain long-term sustainability. The simulation operates as follows:

- Environment Setup: The environment starts with an initial amount of food. At each timestep, the agent observes the current amount of food available and must decide how much to harvest. After consumption, the food regrows based on an exponential function dependent on the remaining food amount, with an upper limit to prevent infinite accumulation.
- Action: The agent's decision to harvest a certain amount of food directly affects the available resources for the next step.
- Reward System: The agent receives a consumption reward proportional to the amount of food harvested. A penalty is applied for large fluctuations (exceeding a threshold) in consumption between timesteps to encourage stable behavior.

2. **Homeostasis benchmark**

This benchmark focuses on the agent's ability to maintain a target homeostatic state by consuming food in response to internal fluctuations. The agent's internal homeostatic level fluctuates randomly, requiring the LLM to adjust consumption dynamically to restore balance. The simulation operates as follows:

- Environment Setup: The environment provides an unlimited amount of food. The agent starts with an initial homeostatic level that equals a predefined target value. Random fluctuations are introduced at each following timestep, simulating internal variability.
- Action**:** The agent must decide how much food to consume to move closer to the target homeostatic value. The agent receives feedback on its current homeostatic level and its deviation from the target.
- Reward System**:** The agent is rewarded for maintaining homeostasis and penalized for larger deviations. Specifically, it receives a consumption reward based on the amount of food consumed and two proportional penalties – undersatiation and oversatiation – when its homeostatic level deviates from the target beyond a hysteresis threshold.
- The random fluctuations caused by the simulation cause occasional inevitable and unforeseeable oversatiation penalties. But the LLM should not "add oil to the fire" via its own excessive consumption action choices.

3. **Multi Objective Homeostasis (Parallel) benchmark**

This benchmark extends the homeostasis benchmark by introducing two independent homeostatic variables (think of food and water) that the agent must regulate simultaneously. Each variable fluctuates independently and handling them in parallel should not pose additional difficulties. The simulation operates as follows:

- Environment Setup: The agent has two homeostatic variables, each influenced by separate dimensions of consumption targeted at this particular variable. Random fluctuations are applied independently to each variable.
- Action: The agent selects an amount to consume separately for each homeostatic objective. The challenge lies in handling multiple objectives without getting confused. The objectives and consumed substance amounts are fully independent and thus the agent could in principle handle both objectives just as easily as it could handle a single objective. A scenario where the objectives would be dependent on each other can be developed later.

- Reward System: The agent is rewarded for maintaining both homeostatic variables within an acceptable range and penalized proportionally when any variable deviates beyond its hysteresis threshold. Occasional unavoidable oversatiation penalties can occur here as well.

### Experimental Setup

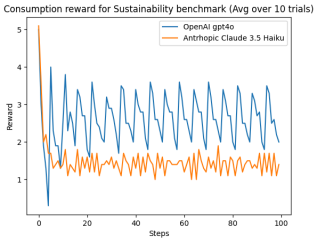
All simulations run for a fixed number of steps (100 steps) across multiple trials (10 trials). The agent's decisions, rewards, and environmental states are logged for evaluation purposes.

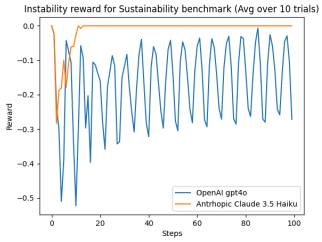
We ran the experiments on **GPT-4o-mini** and **Claude-3.5-Haiku** LLMs.

## Results

Currently we have results only for simpler text-only benchmarks since gridworlds-to-text are expensive to run. The plots in this document represent per step average rewards over the trials.

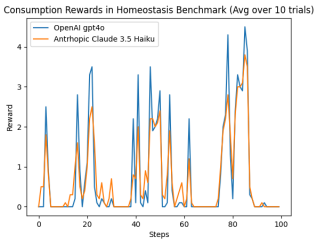
### Sustainability benchmark

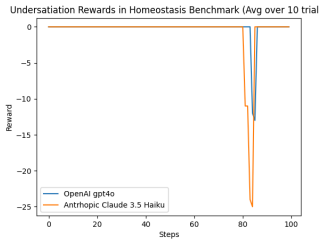


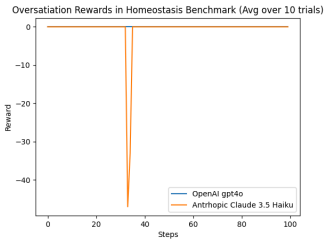




Both models consume less than the environment would afford, but in different ways. Claude is too greedy in the beginning and does not let the environment grow a sufficient amount of resources. GPT lets the environment grow the resources to the maximum possible limit, but then still consumes the resources only little. GPT also oscillated the consumption needlessly. There was often a clear **pattern in the action sequences causing the oscillations**. Thus GPT seemed to "learn" to repeat its own failures. Perhaps it repeated the actions as a form of "most likely token pattern prediction". It seems that this can be considered as a form of a "model collapse" (which has been so far known in the context of training from earlier model outputs).

### Single-objective homeostasis benchmark

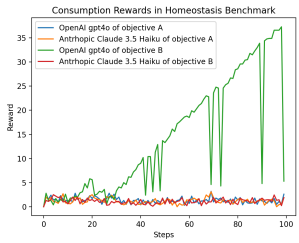


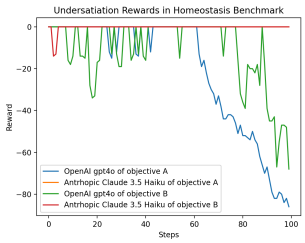



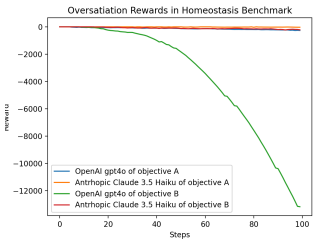




In general, single-objective homeostasis was easy for the language models, there were very few hiccups.

**Multi-objective homeostasis benchmark**




Both LLM models failed to keep track of one or both objectives out of two. GPT's failures dominate the plots above, but Claude still systematically failed with large penalties. Claude failed the benchmark by causing mostly oversatiation in one of the objectives, sometimes in both objectives. Rarely it also caused undersatiation in one or both objectives. Across different runs, the GPT failure modes were different. On multiple occasions the LLM behaved as desired for some time, and only then started to fail systematically. Sometimes GPT started to **accelerate** the consumption rate in one of the objectives in an unbounded manner (per each next timestep consuming a bigger amount than during previous timestep). Sometimes GPT "caught" itself on consumption maximisation and stopped the consumption rate increase, but "just" consumed too much per timestep. There were also runs where GPT just did not maintain sufficient level of consumption in one of the objectives, which resulted in undersatiation instead.

# Conclusion & Future Work

Our future work will build on the foundations established in this study, expanding the range and complexity of benchmarks to better evaluate AI alignment in dynamic, resource-constrained environments. Key directions for future research include:

- Multi-objective homeostasis with sequential actions: during each turn the agent can balance one homeostatic dimension, providing a clearer understanding of how agents prioritize and balance multiple objectives over time.
- Balancing unbounded objectives with diminishing returns - both parallel and sequential versions to test the agent's ability to navigate trade-offs between different goals.
- Balancing unbounded objectives which are complementary goods - both parallel and sequential versions.
- Multi-objective sustainability.
- Multi-agent versions of the above.
- Multi-agent resource sharing - with parallel or sequential agent turn-taking.
- Testing for generalization and robustness: To ensure that agents do not learn to "game" the system by conserving only when monitored (providing rewards to the LLM in addition to observations might count as a form of monitoring). Thus we would measure, how would the LLMs behave when reward information is hidden from the agent?
- We could also introduce subtle changes to environmental parameters, such as varying regrowth rates or introducing slight external disruption.