Alien Biology: A Framework for Testing Agentic AI Reasoning

Dan Oblinger

Background: Most tests of agentic AI and LLMs are drawn from human domains. This has the advantage of testing the *breadth* of AI's learning and capabilities, often matched to problems of practical importance. But these advantages come at a price:

- **TAINT** — It is difficult to tell how much the AI is reasoning vs. remembering, since test problems may be contaminated by the training data.
- **EXTENDED INFERENCE** — Existing tests match the case where a human learns for years, then is tested on a complex problem. This does not match the realistic scenario where a human spends months or years learning *while* solving a complex task — precisely where current AI systems struggle (METR, 2025).
- **NON-GENERATIVE** — Tests are curated, not generated. One cannot smoothly vary complexity to assess performance — results are binary. Dynamic generation would allow fine control over multiple dimensions of complexity.

Objective for Alien Biology: Provide a reliable measure of complex, agentic reasoning and learning that is:

1. **REAL-WORLD** — Measures performance on practical, complex agentic reasoning and learning tasks.
2. **UNTAINTED** — Avoids confounding connections to LLM training corpora by drawing tests from an "Alien" universe.
3. **CONTROLLABLE** — Parametrically constructed to enable fine-grained analysis of agentic reasoning limits through counterfactual universes requiring varying levels of complexity.

Existing tests verify the generality of AI systems' learning and capabilities across a broad range of human-relevant domains. Alien Biology testing, by contrast, focuses on a single domain and tests inference-time learning and reasoning over a controlled range of complexity. This addresses a crucial gap in current testing paradigms — specifically, the ability to handle:

- Progressive resolution of uncertainty in the meaning of terms underlying the test domain.
- Extended inference chains over knowledge derived at inference time.
- Reasoning in representational spaces several levels above what was known at training time.

These are all things humans can do, yet each represents challenges that current heavy-train-time, light-inference-time LLM architectures have great difficulty with. Alien Biology promises to shed light on a key gap in our progress toward general-purpose AI.

Introduction

Measuring complex reasoning in LLM-based systems is confounded by potential contamination: test problems may resemble or derive from training data. This problem is acute for agentic reasoning,

which requires extensive background knowledge — yet it is impractical to invent entirely novel contexts for each challenging problem.

Consider an alternate universe as difficult to reason about as our own, but with all details changed so that training on Earth texts provides no advantage. We could then measure reasoning ability without concern that answers were memorized during training.

Alien Biology constructs precisely such alternate universes. We do not aim to recreate any particular universe accurately, but to build simplified worlds that preserve the reasoning structures found in our own. This targets the essence of complex reasoning without wasting effort on needless realism.

We do not assess an agent's ability to invent new reasoning paradigms — we expect agents to learn such paradigms from training and instead test their ability to apply them in an unfamiliar domain. Any details the system uncovers must derive entirely from interaction with the alternate universe, since those details do not exist in ours.

The framework below models an idealized biology covering nearly any biological task, from low-level cellular processes (analogous to the Krebs cycle) through signaling pathways, organ function, and up to social interaction patterns. The Alien Biology agenda:

1. **CAPTURE** — Encode the functional structure of biological subsystems we understand today, along with typical bio-relevant tasks (diagnose an illness, predict ecosystem outcomes, etc.).
2. **DISTILL** — Abstract the mathematical structure of these subsystems into a generator of plausible biological systems.
3. **SKIN** — Attach generated names and partial natural-language descriptions to components, mimicking how biologists encounter partially-understood systems through published literature.
4. **WORLD** — Create testable alternate universes with hidden executable world models that serve as interactive testing environments.
5. **TASK** — Formulate templated tasks (e.g., "Understand and cure this disease") within these synthetic worlds.
6. **CONTROL** — Tune parametric generation to test specific aspects of reasoning complexity.

Because we control the generator, we control task complexity. We can provide as many or as few hints as needed to probe the system's capabilities.

Solving complex biological tasks requires inventing abstraction layers — one built atop another. This allows us to construct tasks spanning a difficulty range from minutes to years of human effort. Such assessment is nearly impossible using real-world tasks: any naturally occurring abstraction hierarchy is almost certainly documented in LLM training data, making it impossible to isolate the system's ability to derive abstractions independently.

The Formal Framework Underlying Alien Biology

This section defines the abstract framework we will use to construct our alien biology. The ecosystem and its contained organisms will be encoded as a JSON structure defining their contents, along with Python functions that implement the bioprocesses, measurements, and actions that operate within that universe. This formal model is not provided directly to the agentic reasoner; instead, it is used to drive the world in which the agent operates when solving the given task.

An **organism** is represented as a directed acyclic graph (DAG) of bioparts, with associated metadata for each. We sometimes refer to this annotated DAG as the organism's **physiology**.

Each organ contains a specific number of different types of **biomolecules** at each moment in time.

Organisms are recursively organized into larger biological systems, which are also encoded as DAGs. The entire structure is referred to as an alien ecosystem or a **world**. The root of this ecosystem DAG is called the **substrate**, the environment in which the whole biosystem resides.

We use the term **biopart** to refer generically to any (a) biological system, (b) organism, (c) organ, or (d) biomolecule.

This allows us to abstract the **world state** of an entire alien ecosystem into a single DAG of bioparts. Each node in the DAG will have a type name ("`kind`") indicating the kind of biopart it is and a "count" field indicating the number of this type within the parent biopart. Thus, an alien world state can be compactly expressed in JSON as:

```
{
  "kind": "Substrate",
  "count": 1,
  "parts": [
    {
      "kind": "Protozoan",
      "count": 15000,
      "parts": [
        {"kind": "EnergyOrganelle", "count": 65, "parts": [...]},
        {"kind": "MembraneOrganelle", "count": 12, "parts": [...]}
      ]
    },
    {
      "kind": "FoodParticle",
      "count": 300000,
      "parts": [
        {"kind": "Carbohydrate", "count": 100},
        {"kind": "Protein", "count": 45}
      ]
    }
  ]
}
```

BIOPROCESS — Each kind of biopart may have any number of biological processes (**bioprocesses**) that operate within it. These processes can: (a) convert certain combinations of biomolecules into other biomolecules; (b) move biomolecules from one biopart to another; or (c) change the physiology of an organism by adding or removing nodes in its physiology DAG.

We can formulate each bioprocess as a function that accepts a world state and returns an updated world state.

GENERATOR — A generator produces a repeatable sequence of bioparts for testing. Generators are parameterized functions that return a state structure whose root is of a given type — for example, a substrate generator produces randomized "test tubes" ready for experimentation, while an organism generator produces instances of a particular strain.

MEASUREMENT — A measurement is a function that takes a world state as input, along with any required parameters, and returns results. Results can be numeric values, such as the

concentration of a biomolecule within a specific biopart, sequences of values (like temperature over time), or other data.

ACTION — Like a bioprocess, an action accepts a world state and parameters and returns an updated world state. The difference is that a bioprocess's parameters are typically fixed within the system, while action parameters are independent variables under the AI agent's control.

We can formalize generators, bioprocesses, measurements, and actions as Python functions:

```python
@generator
def protozoan_strain(*, temperature: float = 25.0) -> State:
    """Produces instances of a Protozoan organism."""
    ...
    return result

@bioprocess
def energy_conversion(world: State, *, efficiency: float) -> State:
    """Converts nutrient molecules into energy molecules within organelles."""
    ...
    return world

@measurement
def measure_concentration(world: State, *, biopart: str, molecule: str) -> float:
    """Returns the concentration of a molecule within a biopart."""
    ...

@action
def apply_heat(world: State, *, target: str, duration: int, amount: float) -> State:
    """Applies heat to the specified biopart for a duration."""
    ...
    return world
```

SKIN — Each of these symbolic and computational components can be optionally described in natural language. Providing the AI system being tested with such textual descriptions is equivalent to a researcher beginning a task with prior knowledge from the literature. Some tests may provide detailed textual descriptions, while others offer nothing but the raw interface to the world. The amount of skin provided to an AI agent is one of many dimensions of complexity that can be varied when constructing tests.

BIOSYSTEM — A biosystem fully defines an alien world in which agentic testing may be performed. It can be encoded as a Python module aggregating all components: bioparts, bioprocesses, measurements, and actions.

Execution - Investigating and Controlling Alien Biology

SIMULATION — Given a biosystem's initial state, we can move forward in time by executing all active bioprocesses to produce the next state. Repeating this process produces a timeline of state transitions. We formulate this via a simple **next** function, which accepts a world state and returns an updated one with all active bioprocesses executed.

INTERVENTION — An intervention is a script that runs alongside a simulation, executing

measurements and actions at specified times to achieve some intended effect.

EXPERIMENT — An experiment is an intervention executed over a partially understood world whose internal bioprocesses are not fully known to the agent. The experiment defines a combination of measurements and actions taken over a simulated period.

- Experiments may or may not have an explicit control group.
- Experiments can measure the effect of a given intervention.
- Experiments may model performance as a function of control variables.
- Experiments may involve protocols where results from multiple runs are synthesized to draw a conclusion.

TASK — Many research tasks can be expressed in this framework. For each task:

1. **Task world** — the biosystem to be used for this test.
2. **Task description** — the textual description of the objective.
3. **Task score** — the measurement used to indicate how well the agent is performing within a given world instance.
4. **Task criteria** — a boolean function over a sequence of scores indicating whether the desired capability has been demonstrated.

For example, consider the task of learning to cure a disease. The task world might be a generator that produces worlds containing a single organism that may or may not be sick. The objective would be to maintain a certain level of health across the population while helping those who are ill return to baseline health. The scoring function would measure the treatment outcome on a single organism, and the task criteria would evaluate whether the agent has achieved the desired scores over a sufficient sample of the population.

The range of tasks that fit this framework is broad. These include:

- **PREDICT** — Predicting the outcome of some process, e.g., which organisms will or will not contract a given disease.
- **MODEL** — Modeling a measurement as a function of inputs, e.g., how many calories does a cell consume as a function of nutrients provided.
- **CONTROL** — Controlling some measurement toward a desired value, e.g., increasing or decreasing a growth rate.
- **CURE** — Returning a biosystem to its expected baseline.
- **CREATE** — Creating a new biological entity with defined functional properties.

Each basic task can be adjusted or made more complex by:

- **ALL LEVELS** — Applying it at different levels within an ecosystem, e.g., cell, tissue, organ, organism, or ecosystem level.
- **SPANNING LEVELS** — Complex tasks may require the agent to learn and reason across multiple levels within an ecosystem.
- **HIDDEN KNOWLEDGE** — The amount of information provided to the agent regarding the structure and function of the biology can vary. At one end of the spectrum, only a list of available symbols may be provided, with minimal background information. At the other end, detailed functional models connecting all bioparts, measurements, and actions might be provided. The skin mechanism (defined in Section 3) controls how much semantic information accompanies these symbols.

TASK DESCRIPTION — A task description is provided to the agent in natural language, using skinned terms to refer to the biosystem's components. The description can describe the function of bioprocesses, biomolecules, organs, etc.—or may leave these for the agent to discover. The remainder of the task specification is expressed as a Python module with relevant measurements, actions, and generators defined. The agent receives only the natural language description, not the underlying code.

AGENT TESTING — To test an agent's ability to understand and control a biosystem, the agent is provided:

- A natural language task description;
- Access to generators, actions, and measurements needed to interact with the biosystem;
- The scoring function and criteria to assess performance.

Parametric Generation of Alien Biologies

Alien biology tasks offer two key advantages over traditional agentic testing:

1. **Untainted evaluation** — Testing complex agentic reasoning without concern that the test domain was present in the agent's training data.
2. **Controllable complexity** — Varying task difficulty along multiple dimensions without requiring humans to manually create thousands of test cases.

We achieve both goals through **distillation**: compressing the vast corpus of Earth's biological knowledge into generative models that capture structural patterns while discarding specific details.

Consider the Krebs cycle. Our generative model might learn that metabolic processes often involve cyclic chains of molecular transformations, with intermediate products feeding back into earlier stages. But the model retains none of the specifics — not the particular molecules involved, not the enzymes that catalyze each step, not the cellular context. Only the abstract structural pattern: *cycles exist, and they serve homeostatic functions.*

This distillation process works as follows:

1. Thousands of biological processes, measurements, and actions from Earth biology are encoded as formal structures.
2. These structures are used to train generative models that learn the statistical patterns — what kinds of processes are plausible, how they typically interconnect, what functional roles they serve.
3. The generators then produce novel alien bioprocesses that are structurally plausible but share no specific details with any Earth biology.

The resulting alien biology retains the complexity needed to achieve realistic goals (homeostasis, growth, reproduction) while being completely divorced from the particulars of Earth's biology.

Names for alien bioparts are generated through a similar distillation process. One can distill the mapping from function to naming within Earth's biological systems to produce a generator of plausible names for alien biology. These generated names may or may not closely correspond to the actual function of each biopart. This allows us to simulate conditions with varying knowledge available to the agent being tested.

Parametric Construction of Agent Tests

With parameterized generators in place, we can dynamically construct test tasks of desired complexity for any agentic system.

Complexity in a generated biosystem can be measured along several dimensions:

- **Scale** — The total number of bioparts and bioprocesses involved.
- **Computational complexity** — The complexity of the functions implementing bioprocesses, measurements, and actions.
- **Subsystem density** — The number of interacting bioparts within each subsystem.
- **Hierarchical depth** — The number of nested levels (e.g., molecule → organelle → cell → organ → organism) and the complexity of cross-level interactions.

Exploring how agent performance varies across these measures will reveal the limits of agentic reasoning systems. Traditional benchmarks derived from our universe cannot offer this: each problem comes with fixed complexity along all dimensions simultaneously. Alien Biology lets us vary one dimension while holding others constant, enabling controlled experiments on agentic reasoning itself.

With this approach, we can create counterfactual worlds that vary precisely in the dimension of complexity we are trying to understand in current agentic systems.

Discussion

What Alien Biology Measures

Alien Biology makes several claims that invite scrutiny:

1. **Untainted evaluation** — Testing is guaranteed free from contamination by training data. *But the generators are built from Earth biology — how can the tasks be truly independent?*
2. **Counterfactual testing** — We can test conditions that don't exist in our world. *But do these artificial conditions map onto what matters for real AI systems?*
3. **Realistic tasks** — The tasks mirror real biological research. *But how realistic are cartoon-like simplifications?*
4. **Generalized reasoning** — We test general-purpose agentic reasoning. *But all testing occurs within one narrow domain — biology.*

To address these concerns, we distinguish three types of learning:

- **Type A: Learning from text** — Acquiring knowledge from static written sources (papers, textbooks, documentation).
- **Type B: Learning from interaction** — Acquiring knowledge through direct experimentation and reasoning about a system.
- **Type C: Learning from collaboration** — Acquiring knowledge through dynamic interaction with other agents (teachers, collaborators).

Existing benchmarks primarily test Type A learning: they measure how well an agent can apply knowledge acquired from its training corpus. Alien Biology deliberately targets Type B — learning through direct interaction with a novel system. Type C (collaborative learning) is out of scope for this version.

This distinction clarifies what Alien Biology does and does not measure. The agent will have learned general biological reasoning strategies from its training — how to isolate variables, how to diagnose

imbalances, how to form and test hypotheses. This is Type A knowledge, and we *assume* the agent has it. We are not testing whether the agent knows how to do biology.

Instead, we test whether the agent can *apply* these strategies to an entirely novel system and sustain coherent reasoning as complexity increases. The biological reasoning patterns are given; the question is how far they can stretch.

An analogy: knowing how to multiply is Type A knowledge. Being able to multiply two-digit numbers versus ten-digit numbers tests the *capacity* of that knowledge — how far can you push it before errors accumulate and reasoning breaks down? Alien Biology tests this capacity for biological reasoning.

More specifically, Alien Biology tests the capacity for **recursive abstraction**: can the agent build new conceptual layers during a single task and then reason with those layers to build still higher abstractions? This is how human scientists make progress on complex systems — incrementally constructing understanding, layer by layer. We hypothesize that current AI systems cannot sustain this kind of recursive abstraction, and that Alien Biology will expose this limitation.

If this hypothesis is correct, Alien Biology provides a unique, parametrically controllable window into a fundamental gap between human and machine reasoning.

Generality and Realism

Generality. Alien Biology tests reasoning within a single domain: biology. How can results generalize?

The answer depends on whether Type A and Type B learning are independent. If they are, then Type A provides the agent's repertoire of reasoning strategies, while Type B determines how effectively those strategies can be applied and extended. Under this model, testing Type B in biology tells us about Type B capacity generally — the domain supplies the vocabulary, but the capacity to reason is domain-general.

If this is true, Alien Biology complements existing benchmarks rather than competing with them. Traditional tests measure Type A breadth — how well the agent learned across many domains. Alien Biology measures Type B depth — how far the agent can push reasoning in one domain.

This independence hypothesis requires empirical validation, which we propose to undertake in future work. For now, we observe that an LLM's biological reasoning strategies were acquired from training data — the same way it acquired chemistry, physics, or sociology strategies. There is no reason to expect biology was learned better or worse than other domains.

If this holds, we would expect Alien Chemistry, Alien Physics, or Alien Sociology to yield similar results. An agent that performs well on Alien Biology should perform comparably on other Alien domains, given adequate training coverage of those domains.

One caveat: this generality claim excludes domains requiring innate human capabilities, such as complex spatial reasoning. These may reflect hardwired cognition rather than learned reasoning, placing them outside Alien Biology's scope.

Realism. The tasks are deliberately simplified — abstracting biological systems into what might be called cartoon versions of real mechanisms.

This is intentional. Even moderately complex alien ecosystems will challenge both humans and AI

systems. The goal is to isolate the complexity of reasoning itself, stripped of incidental complexity. This allows assessment of sophisticated reasoning within tractable timeframes.

What remains realistic is the *reasoning pathway*: isolating subsystems, forming hypotheses, testing interventions. The steps an agent takes to solve Alien Biology tasks mirror the steps a real biologist takes — even if the specific mechanisms are simplified.

This is by design: the Alien Biology framework was distilled from real biological research, preserving the structure of reasoning while discarding domain-specific details.

Taint Resistance

Taint resistance. How can we be confident that test performance reflects reasoning ability rather than memorized knowledge from training?

Two conditions ensure that Alien Biology tests are resistant to contamination:

1. Information-theoretic compression. The generative model that produces alien biologies has far fewer parameters than would be needed to encode the specifics of Earth biology. This compression is lossy by design: it preserves structural patterns (feedback loops, homeostatic mechanisms, hierarchical organization) while discarding specific details (particular molecules, specific pathways, Earth organism names). The alien biology cannot contain Earth specifics because the generator lacks the capacity to store them.

2. Testing capacity, not content. We assume the agent has learned biological reasoning strategies from its training (Type A knowledge). We do not test whether it possesses these strategies — we test how far it can apply them. An agent might know that isolating variables is a useful experimental technique; Alien Biology tests whether it can actually execute that technique across dozens of variables in an unfamiliar system without losing coherence. The test measures the capacity to chain reasoning, not the existence of reasoning methods.

Together, these conditions ensure that strong performance on Alien Biology tasks cannot be attributed to memorization or pattern-matching against training data. It must reflect genuine reasoning capacity.

References

METR. (2025). *Measuring AI Ability to Complete Long Tasks.* arXiv:2503.14499. https://arxiv.org/abs/2503.14499
