AlienBiologyWhitepaper

Background: Most tests of agentic AI and LLMs are drawn from various human areas of activity. This has the advantage of testing the *breadth* of AI's learning/capabilities that are often carefully matched to problems of *practical import*. But these advantages generally come at a price:

- **TAINT**—It is often difficult to tell how much the AI is reasoning vs. remembering, since nearly all problems may be tainted by their relationship to materials in the training sets.
- **EXTENDED INFERENCE**—Existing tests match the case where a human learns for years, then is tested on a small, hard problem. But do not match the realistic scenario where a human spends months or years *learning* while solving a complex task, yet this is precisely where current AI systems fail. (see METR results).
- **NON-GENERATIVE**—These are individually curated test problems, not mechanically generated. Thus, one cannot smoothly vary their complexity along various dimensions to assess small changes in system performance; it either solves the problem or does not. Dynamic generation would allow fine control over multiple dimensions of problem complexity.

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

1. ***REAL-WORLD*** - measures performance on practical, complex, real-world-relevant agentic reasoning/learning tasks.
2. ***UNTAINTED*** - avoids confounding connections to LLM training corpora by drawing tests from an "Alien" universe.
3. ***CONTROLLABLE*** - is parametrically constructed in ways that allow fine-grained analysis of the limits of agentic reasoning by creating counterfactual universes, each requiring varying levels of inferential 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 of human problem solving and tests inference time, learning, and reasoning over a controllable range of complexity. This covers a crucial gap in current testing paradigms. Specifically, inference-time testing of the ability to handle:

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

These are all things humans can do, yet each represents hurdles that the current heavy-train-time, light-inference-time LLM architectures have great difficulty with. Thus, Alien Biology promises to shine a light on a key gap in our progression towards true AGI.

Introduction

Measuring the ability of today's LLM-based systems to perform complex reasoning is often confounded by potential contamination of their training data by problems related to those one is using in the testing. This problem is especially acute when measuring complex or agentic reasoning since both require complex background knowledge. It's not practical to invent entirely novel contexts for each challenge problem.

If only there were some alternate universe that was as difficult to reason about as our universe, but with all details changed such that training on text from our universe afforded no advantage in answering detailed questions drawn from this alternate universe. Then, we could measure an agent's ability to reason about this alternate universe w/o concern that knowledge about the questions we are asking was somehow provided during the LLM construction.

The Alien Biology framework described below is designed to allow us to construct just such alternate universes for testing our agentic systems. We aim not to recreate an accurate model of any particular universe but to build new universes simplified to target reasoning structures similar to those in our own world. This gets to the crux of the complex reasoning w/o wasting effort with needless realism.

We do not aim to assess the agentic system's ability to invent new reasoning paradigms. Only a few humans throughout history have accomplished such a feat. Indeed, we expect the agentic system to learn relevant reasoning paradigms from its training corpora and instead test its ability to apply them to an alternate universe where all the details have been changed. This provides certainty that any details the system uncovers must have been derived entirely from its interaction with the alternate universe, since none of them even exist within our universe.

Below is an idealized model of biology that we believe covers (in a simplified way) nearly any task one might undertake relative to nearly any biological. This framework can encode low-level functioning within a cell, like the Krebs cycle, signaling pathways coordinating groups of cells, the functioning of whole organs like the liver, and all the way up to the highest level of interaction patterns found between socially interacting animals.

The Alien Biology agenda is to:

1. CAPTURE: Capture the functional structure for the many bio subsystems that we understand at all levels of biology today, as well as the range of bio-relevant tasks that we consider today. (e. g. cure an illness, predict ecosystem or cell outcome, etc.)
2. DISTILL: Use a diffusion or other model to abstract the mathematical structure of these many subsystem models into a generator of plausible functioning biological subsystems.
3. SKIN: "Skin" those functional systems by attaching diffusion-generated namings for relevant components and partially-explained functional descriptions of these generated systems in natural language, just as a biologist today might partially understand underlying biological processes from published background papers.
4. WORLD: Once an alien biology and chemistry are created and textually described, they form a test bed with a hidden executable world model that can be used as an interactive testing environment for testing agentic reasoning over complex novel tasks.
5. TASK: Templated test tasks like "Understand and cure this disease" can be formulated within these synthetic worlds.
6. CONTROL: By controlling the parametric generation of these worlds and tasks, one can finely tune one's testing of various aspects of the complexity of the learning/reasoning task.

Because we control the generator for these ecosystems, we can control the complexity of the learning/inference tasks we generate. We can provide the AI system with as many or as few hints as we choose to test its ability to solve alien puzzles.

As with real biology, solving the more complex versions of these tasks forces the reasoner to invent new abstraction layers one on top of another to address the overall task in question. This allows us to construct tasks that might take a biologist 5 minutes or 5 years to complete. This kind of assessment, in particular, is nearly impossible to test in an untainted way using native tasks; any naturally occurring hierarchy of abstractions is almost certainly to have been well documented within the text used to train the LLM, making it impossible to assess the system's ability to derive and use those abstractions.

The Formal Framework Underlying Alien Biology

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

An **organism** is represented as a DAG (directed acyclic graph) of **organs** 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 whole structure is called an alien ecosystem or a world. The root of this ecosystem DAG is called the substrate, the environment in which the entire 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 ("`num`") field indicating the number of this type within the parent biopart within the DAG. Thus, an alien world state can be compactly expressed in JSON as:

```
{"kind": "Substrate", "num": 1,
    {"kind": "Protozoan3", "num": 15000, parts: [{"kind": "energyorganelle4",
"num": 65, ...}, {...}]},
    {"kind": "Food5", "num": 300000, parts: [{"kind": "biomolecule_carb", "num":
100}, {...}]},
    ...
}
```

```
{
    "biomolecules": ["Glycogen," "H20", "Insulin," "Lyucine", "Hemoglobin"],
    "organelles": ["Mitocondria"],
    "cells": ["Red Blood Cell", "Smooth Muscle Cell"],
    "system": ["Lymphatic," "Vasculature"],
    "organ": ["Liver", "Lung"],
    "organism": ["Amoeba", "Cricket"],
    "ecosystem": ["Fresh Water Pond"]
}
```

BIOPROCESS - Each kind of biopart may have any number of biological processes (**bioprocesses**) that operate within it. These processes will:
(a) convert certain combinations of biomolecules into other biomolecules
(b) move biomolecules from one organ to another, or
(c) change the physiology of the organism by adding or removing edges (cells, organs, etc.) in its physiology DAG

Typically, the rate at which each biological process executes is called its efficiency, and its rate formula stochastically controls it at each moment. But more generally, we can formulate each bioprocess as one that accepts a world state and returns an updated world state for the whole ecosystem.

GENERATOR or **BIOSTOCK** - To make observations, run experiments, etc., we need generators that will produce a repeatable sequence of substrates, organisms, etc., for testing. Each of these generators is called a **biostock** or a **generator**; these are parameterized functions that return a state structure whose root is of a given type or set of types. So, a substrate is a parameterized biostock generator that produces a sequence of randomized "test tubes" ready for testing. An organism biostock is a parameterized generator that produces a sequence of instances of a given strain of some organism type.

MEASUREMENTS - Of course, one needs to be able to take **measurements** of a biological system in order to study it. In our case, this is simply a function that takes in a world state as input, along with any parameterization required by the measurement, and then returns the measurement's results. The results might be a numeric value like the concentration of a biomolecule within a type of cell, a sequence of

values (like temperature over time), or other data output. The measurement's inputs might include a specification of what biopart of the system one is measuring.

ACTION—The AI agent also needs the ability to act on the Alien Biology in some way. This is accomplished via actions. Like a bioprocess, an action accepts a world state as input along with any parameters required to fully specify the action and returns an updated world state. The difference is that the parameters for a bioprocess are typically fixed within the subsystems in which they operate. Meanwhile, actions and action parameters are typically independent variables under the AI agent's control.

We can formalize samples, biological processes, measurements, and actions as Python functions, as shown here:

```
@generator
def my_organism_strain(*, ...) -> State:
    ...
    assert result["kind"] == "MyOrganism"
    return result


@bioprocess
def adp2atp(world: State, *, ...) -> State:
    ...


@measurement
def measure_concentration(world: State, *, biopart: str, biomolecule: str) -> int
    """Returns the number of a given biomolecule within a given (or all) named
parts of an ecosystem"""


@action
def apply_heat(world: State, *, biopart: str, duration: int, amount: int) -> State:
    """Applies a given level of heat to the specified biopart for the specified
number of time steps."""


class CmdKind(Enum):
    ADD, ACT, MEASURE = "Add", "Act", "Measure"

class Recipe:
    class Step:
        kind: CmdKind
        name:
        start: int
        duration: int
        ...
```

SKIN — Each of these symbolic and computational components can optionally be described textually. Providing the AI system being tested with such textual descriptions of the bio system is equivalent to a researcher beginning a task with knowledge gleaned by others before them. Some tests can provide significant textual descriptions, while others provide access to the world to be understood/controlled, and nothing else. The amount of 'skin' provided to an AI agent is one of many dimensions of complexity that can be varied in testing its capabilities.

BIOSYSTEM — A biosystem fully defines an alien world in which agentic testing may be performed. As we see above, it can be encoded as a Python module, and it is an aggregation of the aspects defined above: biomolecules, bioparts, bioprocesses, measurements, and actions.

Execution - Investigating and Controlling Alien Biology

SIMULATION — Given our biological system's initial state, we can move forward in time by executing all active bioprocesses in random order to produce the "next" state. Repeating this process can produce a timeline of plausible state transitions of that initial world. We formulate this via a simple "**next**" function, which accepts a world state and returns an updated one with all active bioprocesses run.

INTERVENTION—An intervention is a goal-directed script that runs alongside a simulation. It performs measurements and actions to achieve its intended effect.

EXPERIMENT—Using this framework, we can describe an experiment as an intervention script run over a partially understood world composed of an unknown number of bioprocesses. The script describes a combination of measurements and actions taken over simulated time.

- Experiments may or may not have an explicit control group.
- Experiments may measure a given intervention's effect.
- Experiments may model some aspects of performance as a function of their control variables.
- Experiments may involve protocols where results from multiple runs are synthesized to produce a result or draw a conclusion.

TASK - Analogs for many tasks that bio-researchers might take on can be succinctly expressed using this framework. For each task:

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

For example, consider how one might frame the task of learning how to cure a disease. The task setup might be a generator of the world, each containing a single organism that may or may not be sick. The task would be to maintain some measure of health across the population while bringing the sick ones back to some baseline behavior. The scoring function would measure the treatment outcome on a single organism provided by the setup generator, and the overall task criteria would be a boolean that measures if one has achieved the desired scores over a sufficient sample of the population.

The range of biologically plausible tasks naturally fitting into this framing is quite broad. These would include the task of:

- PREDICT - Determining how to predict the outcome of some process. e. g. Which organisms will or will not get a given disease.
- MODEL - Modeling any desired measurement from a system. e.g., How many calories will a cell consume as a function of nutrients provided?
- CONTROL - Controlling some measurement of the system toward some desired value. e.g., increasing or decreasing a growth rate.
- CURE - Adjusting some biological systems to their expected (defined) baseline.
- CREATE - Creating a new biological entity with some defined functional properties.

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

- ALL LEVELS - Applying it at different levels within a given ecosystem, e.g., cell level, cell group, organ, organism, ecosystem, etc.
- 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 systems of alien biology can vary the tasks. At one end of the spectrum, a simple

list of actions and measurements might be available as a list of symbols with a little background about them, and at the other end of the spectrum, detailed functional models connecting all of the bioparts, measurements, and actions might be provided. Each of these has been generated as a JSON structure and Python function, so it is easy to supply portions of this information and natural text as hints for the agent being tested. (See skinning in the next section for a discussion about natural text.)

SKINNING—Up to this point, the components of alien biology have been described as symbolic or mathematical structures expressed as JSON expressions and Python functions. This structure can be "skinned" by inventing a biological name for each of the bioparts of the system (the biomolecules, bioprocesses, organs, organisms, systems, and substrates of the system). As an aid in understanding the alien system, we might adopt biologically plausible analogs for the functioning of many of the bioparts if we want to simulate a condition where much is understood about the functioning of the various bioparts. Alternatively, just as biologists often do early on, when little is understood about a system, one can adopt generic names that give away little about the underlying functioning of the system in order to simulate the situation when a new biopart has been discovered but little is known about that biopart. Either way, a **skin** is simply a mapping from each biopart identifier onto a name string that is used for describing that biopart in the task specifications below.

TASK DESCRIPTION - Using this skinning of the logical model, one can provide a **task description** in natural language using skinned terms to refer to the biology of the system. This description may or may not describe the function of the bioprocesses, molecules, organs, etc., or these bioparts to be solved by that agent trying to solve the task. The remainder of the task specification is expressed as a Python module with relevant measurements, actions, and generators defined. We assume the contents of this model are not available to the agent being tested; the only information they have is the natural language task description.

AGENT TESTING - Tying all of this together we can test an agent's ability to understand and control these biological systems. They are provided:

- A natural language task description.
- Task setup, actions, and measurements are required to simulate, investigate, and control the biosystem.
- And the scoring measures and outcome criteria needed to assess performance.

Parametric Generation of Alien Biologies

The promise of alien biology tasks over traditional agentic testing is the ability to:

1. Test complex agentic learning and reasoning in realistic scenarios without concern that the bioparts of the task being learned were somehow already provided during the training of the LLM agent model itself.
2. Control various dimensions of task complexity in a fine-grained, controlled way without requiring humans to create tens or hundreds of thousands to test tasks.

We achieve both of these goals by dynamically generating new alien biologies and generating tasks within those biologies. We can use the hundreds of thousands of understood processes, molecules, organs, and systems in order to create realistic universes and realistic tasks within that universe. But we dramatically compress these hundreds of thousands of processes, molecules, organs, etc., into quite a small model for generation. Thus, the nature of biological processes is retained. e. the idea of a cycle of molecules as found in the Kreb cycle without any specific information about the specifics of that cycle, which it exists within a system, what its purpose might be, etc., just the bare structural information with work from. Thousands of measurements, actions, and processes are expressed in Python (using LLM models), and then again, these structures are distilled into generators of plausible but quite randomized processes. The system retains the structure required to achieve some homeostatic or other goal but is divorced from any of the particulars found within Earth's biology.

Model skin names are constructed in a similar fashion. One can distill the mapping from formal function onto naming within Earth's biological systems in order to produce a generator of plausible naming for the alien biology, which may or may not tie closely to the functioning of each biopart within that biology. Thus allowing us to simulate conditions of varying background knowledge available to the agent being tested.

Parametric Construction of Agent Tests

Once distilled, parameterized generators for alien biology have been constructed; we can use them to dynamically construct test tasks of the desired complexity for our agentic system.

There are many plausible measures of complexity for our generated biology:

- Total number of bioparts, processes, molecules, organs, etc involved.
- The number of lines and operators in the Python code used in processes, measurements, and actions.
- The complexity of the interacting bioparts within the dynamically constructed sub-systems of the larger system
- The number and complexity of the interactions between the layers of the full system.

Ultimately, exploring how agent performance varies across these different measures will give us an understanding of the limits of our agentic reasoning systems, which we cannot access today since we pragmatically have no way to vary these aspects of the problems we laboriously obtain from our actual universe.

By distilling the working of our biological universe into a generator, we are able to create counterfactual worlds that vary on precisely that aspect of complexity we are trying to understand/improve within our current agentic systems.

Discussion

Alien Biology is a unique approach for agentic testing. We make interesting promises about it, which leads to natural questions regarding those promises. It allows:

1. Testing that is guaranteed to be untainted by any information provided during the construction of the agentic LLM system.
   However, given that the distilled generators are built from bio-data, in what way are these tasks independent?*
2. Testing of counterfactual conditions that do not exist within our world.
   But can these map onto the key aspects of our agentic systems we hope to test?
3. Testing on realistic tasks
   But how realistic are these tasks, and in what ways are they, and are they not realistic?
4. Testing of generalized ability to perform complex agentic reasoning.
   But how generalized is this testing? It all occurs with a tightly defined representation of biology and biological tasks.

In answering these questions, we split the task of general-purpose learning and inference crudely into three parts:
(A) Obtaining knowledge from distilled into static (written) forms, other agents in the world.
(B) Obtaining knowledge in isolation via thinking or interacting with the world.
(C) Obtaining knowledge from dynamic interaction in the world along with other possibly collaborating (teaching) agents.

We believe the first kind of learning and inference is well measured by existing agentic testing. In this case, we *want* the agent to be exposed to static written forms and then measure the performance that results. Alien biology is not trying to test type A. Indeed, we hope to be pretty isolated from type A

knowledge and instead are focused on type B learning and inference. Type C is out of scope for this first version of Alien Biology.

Splitting the learning/inference task this way allows us to consider these isolation and generality claims better.
Alien biology is about ***biology***! Thus, any generalized knowledge about how processes connect and how one might proceed in testing or understanding such a system is available within the training provided in constructing an agentic LLM. Thus, assessing these systems on how well they learned, for example, to isolate the functioning of a biomolecule or its strategic approach to diagnosing and correcting an imbalance within a system, would make no sense. Such knowledge was available during agent construction. But notice, all of this knowledge is of type A above. It is not the kind of knowledge that Alien Biology is designed to measure; indeed, much of that knowledge is implicitly embedded in the framing of Alien Biology itself.

Instead, we take that structure as a given and assess how well the agent can apply and reason with that knowledge over ever-increasing structural complexity. It's like the difference between being able to multiply two two-digit numbers and being able to multiply two ten-digit numbers. How complex can one's type-II reasoning get before the agent confuses itself and cannot proceed?

It also measures the degree to which the agent can recursively build new abstractions while solving a single problem and then use them layer by layer within that task. We have ample evidence that current agentic AI is incapable of this generality, while humans are. Humans can repeatedly apply biological strategies to incrementally uncover the functioning of even quite complex, multi-layered alien systems, while (we believe) current agentic AI systems will not.

If true, it provides a unique and parametrically controllable window into the gap between humans and current agentic AI.

But how general is this testing framework? It is all about biology, occurring within a very narrowly defined framing of biology at that. The answer depends upon how independent type A and type B learning/reasoning is above. Suppose they are quite independent of each other. In that case, one can understand type A reasoning as that which provides the raw structures, processes, and strategies the AI agent has to work with, and type B reasoning uses those processes to perform its reasoning/learning. If this is true, measuring type B reasoning/learning in Biology provides a good proxy for how these systems will perform in doing extended type B inference over other domains, while existing type A tests from other researchers provide assessments of the generality of these systems over a range of domains.

We can't be sure about this without testing (which we propose to do with later versions of Alien Biology). For now, we observe that whatever knowledge the AI agent has about biology, it has it because of the training data provided in building its LLM. We have no reason to expect that such training is somehow differentially better in training bio strategies relative to learning/inference strategies for other domains. If this is true, we expect investigating Alien chemistry, physics, geology, sociology, etc., to yield similar results. A system that is good at Alien Biology will be good at Alien Chemistry, too, assuming that the LLM training data contained a good background in chemistry.

(FOOTNOTE: This expected generality will not extend into complex spatial domains or other areas where it seems likely that human agents have special wiring at birth to address these tasks. We view these as out of scope for Alien Biology's measure of learning/reasoning; we only notice that these may not be learned by humans either, but are instead hardwired capabilities.)

So, how realistic are these testing tasks? Well, in one way, they are not very realistic; they abstract complex biological systems into almost cartoon-like simplifications of the actual underlying mechanisms. This is intentional, even with such simplifications in place, the complexity of one of these alien ecosystems constructed with moderate complexity will already be daunting for a person or AI to solve. Our aim with Alien Biology is to capture the fascinating learning/inference complexity in the simplest possible packages. Thus allowing us to assess fairly complex reasoning in tractable periods of time. We believe solving these bio tasks will be composed of very plausible high-level steps of trying to isolate the functioning of individual pieces, etc., so in that way, it has realistic pathways of bio reasoning. But

framed in cartoon-like contexts that simplify the messy details as much as possible, so from a computational and measurement perspective, one can jump right to the "meat" of the task in each case. Thus, the details of alien biology are not so realistic. Still, the approaches one would take in solving these tasks should map well onto the approaches biologists take in solving their problems, as the entire edifice has been constructed by distilling all of that biological work.

How resistant is this testing paradigm to testing "taint" derived from the training materials provided during construction? Given some assumptions about the distillation process and testing tasks, it seems quite resistant. As long as (1) the number of bits allowed in the weight matrix is much smaller than any characterization of Earth biology, we can be sure all traces of Earth specifics in the alien biology are removed. The only thing left is generalizations about things like feedback loops that support homeostasis or other generalized mechanisms. And (2) we only test our agent's ability to string together multiple such strategies, and we are not testing the strategies themselves, so we know its performance is not tied to anything it learned. Instead, it is testing its ability to sustain long chains of learning/inference over what it has learned.

References

- METR - Measuring AI Ability to Complete Long Tasks