Apodex Discovery: Reality Benchmarks and Environments for Evaluating and Building Discoverative Artificial Intelligence

Brian Wang, Bin Feng, Xiaoman Pan, Chenyang An, Felix Liu, Tangqi Fang, Gongbo Sun, Lingfeng Shen, Ning Wang, Handuo Zhang, Feng Chen, Fuchao Yang, Xiang Wang, Jiacheng Lin, Siting Li, Zixuan Liu, Chi Han, Zhenhailong Wang, Kunlun Zhu, Lawrence Zhao, Yueqi Guo, Kailong Wen, Feng Xing, Yiling Guo, Lidong Bing, David Tan, Bo An, Heng Ji, Sheng Wang[†]

Apodex

Apollo did not reach the Moon merely because its engineers could solve difficult equations. It succeeded by turning a distant ambition into a mission architecture of explicit objectives, coordinated systems, simulation, telemetry, verification, and repeated correction. AI now faces a similar transition: frontier models can solve increasingly difficult tasks once the problem, tools, and success criteria have been specified, yet the most consequential real-world challenges rarely arrive in an executable or verifiable form. We introduce **Apodex Discovery**, a framework for building and evaluating *discoverative AI*, referring to AI directed at making genuine discoveries, realized through its operational unit, the *heavy-duty solver*, together with **TRACES**, a reality benchmark we built. Such heavy-duty solvers are complete systems comprising a foundation model, harness, tools, and control policies that pursue extended, stateful, and verifiable investigations.

Apodex Discovery makes three principal contributions toward *discoverative AI*: turning open-ended real-world problems into tractable, verifiable discovery. First, we develop a systematic problem-scouting process that surveyed 561 industries across 16 sectors, assembled 423 high-value real-world problems, and selected 20 for the initial release, spanning retrospective tasks with hidden outcomes and *discovery challenges* evaluated prospectively as new evidence emerges. Second, we formalize *executable environments* through a common environment–task–episode abstraction that provides solvers with the data, tools, constraints, and feedback needed to act, while recording trajectories and verifying key intermediate artifacts and final submissions; these efforts yield **TRACES**, the reality benchmark we released together with this paper. Third, we introduce **HDS6**, a process-verification framework assessing Tools, Repair, Alternatives, Coherence, Evidence, and Scope independently of final-task success. This enables meaningful evaluation even when definitive ground truth is delayed, incomplete, or unavailable.

In adeno-associated virus (AAV) capsid design, Apodex surpassed the task-level published state of the art by 7% across all four tasks: viability prediction, tropism prediction, structure prediction, and generative design. In drug repurposing and reformulation, adding a task-specific biomedical environment improved the mean normalized prediction score of GPT-5.5 and GPT-5.6-sol by 2.5 and 7.6 points, respectively, over the same closed-book backbone. Controlled ablations further demonstrate that the fixed TRACES episode interface enables performance differences to be attributed to specific components of the solver system. By leveraging real-world problems, executable environments, and process evaluation, Apodex Discovery defines a new paradigm that lays a foundation for *discoverative AI*: moving beyond solving predefined benchmarks toward conducting the consequential investigations by which discoveries are made. The TRACES benchmark is available at discovery.apodex.com.

[†] Technical Lead. Corresponding author: sheng@apodex.com.

We gratefully acknowledge Mr. Tianqiao Chen, the project lead of this work, who proposed the concept of the *heavy-duty solver*, defined the six capabilities of the HDS6 process metric, and shaped the overall architecture of Apodex Discovery. We sincerely thank him for the vision, guidance, and support that made this work possible.

Contents

1 Introduction **4**
- 1.1 What is a Heavy-Duty Solver? . . . 5
  - 1.1.1 Problem Manifest . . . 6
  - 1.1.2 Reality-Based Environment . . . 6
  - 1.1.3 Verification Mechanisms . . . 7
  - 1.1.4 Repair Loops . . . 8

2 Scouting High-Value Real-World Problems **9**
- 2.1 Domain Selection . . . 9
- 2.2 Question Collection and Screening . . . 11
- 2.3 Reality-Facing Problems in the Initial Benchmark . . . 12
- 2.4 Discovery Challenges . . . 13

3 Process Verification and Reality-Based Environment **14**
- 3.1 Verification . . . 14
  - 3.1.1 Outcome Verification . . . 15
  - 3.1.2 Process Verification . . . 17
- 3.2 Environments, Tasks, and Episodes . . . 20
- 3.3 Environments . . . 21
  - 3.3.1 Tasks . . . 22
  - 3.3.2 Episodes . . . 23

4 Results on High-Value Problems **23**
- 4.1 Adeno-associated virus capsid design . . . 24
  - 4.1.1 Task Formulation . . . 24
  - 4.1.2 A domain-specific environment improves outcomes . . . 25
  - 4.1.3 Surpassing the published state of the art . . . 25
  - 4.1.4 Results based on process verification . . . 25
  - 4.1.5 Datasets and evaluation . . . 26
- 4.2 Drug Repurposing and Reformulation . . . 27
  - 4.2.1 Task Formulation . . . 27
  - 4.2.2 Scoring Methods . . . 27
  - 4.2.3 Public-100 Benchmark Results . . . 29
  - 4.2.4 HDS6 Trajectory Analysis . . . 29
  - 4.2.5 Repurposing and Reformulation Procedure Design . . . 30
- 4.3 Large language model problems and ablations . . . 31
  - 4.3.1 Solver-configuration ablations . . . 31
  - 4.3.2 Episode-mechanism ablations . . . 34
- 4.4 Verification-Driven Repair . . . 35
  - 4.4.1 Case study: an under-specified clinical safety report . . . 36
  - 4.4.2 Case study: a named estimator that was never run . . . 37

5 Related Work **39**

6 Conclusion **42**

7 Author Contributions and Acknowledgments **42**

A Domains and questions considered by Apodex Discovery **51**

B Adeno-associated virus capsid design full results **54**

C Full harness × model result matrix **58**

D Drug Repurposing and Reformulation Trace Example **68**
D.1 The task as given 68
D.2 Full reasoning trace 70
D.3 Process score 81

1. Introduction

The Apollo program did not succeed merely because its engineers could solve difficult equations. The ambition to "reach the Moon" first had to be transformed into a mission architecture: a precise objective, explicit constraints, coordinated subsystems, simulation environments, telemetry, failure criteria, and repeated cycles of testing and correction. Only within this architecture could individual acts of problem solving accumulate into a consequential real-world achievement.

Artificial intelligence may now be approaching a similar transition. On advanced examinations, mathematical challenges, and software-engineering benchmarks, increasingly capable models can solve difficult problems once the objective, environment, tools, and success criteria have already been defined. Yet humanity's most consequential ambitions, from developing curative therapies for currently untreatable diseases, to achieving commercially viable fusion energy, to discovering new materials with transformative properties, rarely arrive with such an architecture. The next critical challenge for AI may therefore lie not only in building more capable models, but also in constructing the infrastructure that translates their capabilities into solutions to open and consequential problems.

The aim that ultimately motivates this infrastructure is *discovery*: enabling AI systems to reach conclusions that are not yet known, rather than to reproduce answers that already exist. Discovery is the goal, and the question this paper addresses is *how* it is carried out—how an open-ended aspiration becomes a sequence of grounded, verifiable investigative acts that a system can execute and that others can check. We refer to this goal-directed paradigm as *discoverative AI*. Discovery in practice proceeds as a loop, forming a hypothesis, acting on the world, verifying the result, and revising, and automating this discovery loop is the shared ambition of a rapidly expanding body of work, from autonomous scientific agents [1–3] to dedicated efforts to automate the experimental loop of research itself, such as Discovery Loop [4]. What these efforts require, and what remains largely absent, is the infrastructure that makes each turn of the loop executable, verifiable, and improvable.

We identify four components of this missing infrastructure: *problem formulation, reality-based environments, verification mechanisms, and repair loops*. Problem formulation translates an open-ended human ambition into explicit objectives, constraints, assumptions, and criteria for success. A reality-based environment equips the solver with the data, tools, computational resources, experimental interfaces, and human interactions required to act on the problem. Verification determines whether intermediate claims and final outcomes are supported, correct, and relevant to the original objective. Repair loops return the results of computation, experiments, external evaluation, or human judgment to the solver, enabling it to identify failures, revise its hypotheses, and improve its solution. Together, these components transform an under-specified ambition into a tractable, iterative, and evaluable problem that a *heavy-duty solver* can pursue: a solver which is essentially the *operational unit* of discoverative AI.

Moreover, current evaluation paradigms are not designed to measure this form of problem solving. Most benchmarks fall into two broad families. Static question-answering benchmarks, including MMLU [5], GPQA [6], MATH [7], FrontierMath [8], and Humanity's Last Exam [9], evaluate whether a model maps a fixed prompt to a correct answer. They probe important capabilities, but provide no persistent environment, external action, evolving state, or opportunity to respond to intermediate failure. Interactive benchmarks such as SWE-bench [10], WebArena [11], OSWorld [12], and τ-bench [13] move closer to realistic problem solving by placing models in executable environments. Nevertheless, they typically remain bounded to narrowly specified tasks and reduce an episode to a final success signal. They rarely evaluate whether a solver decomposed the problem correctly, grounded its claims in evidence, considered viable alternatives, or

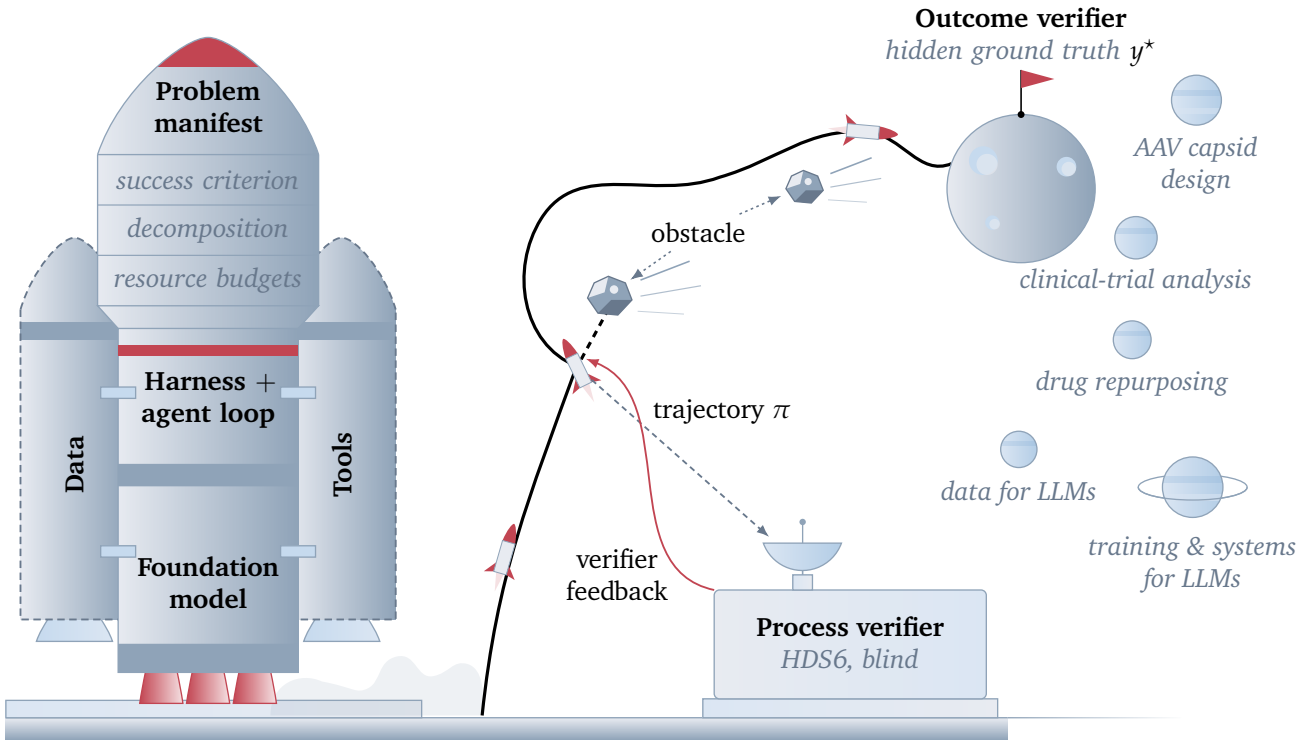

Figure 1: **Apodex Discovery mission architecture for a heavy-duty solver.** The fairing contains the *problem manifest*; the solid core represents the *solver system under evaluation*; and the dashed side boosters form the task-specific *reality-based environment*, which can be exchanged across problems. These components meet at a common *episode interface*. During execution, failed checks trigger verifier feedback and repair, while the full trajectory π is recorded for blind process verification. The final submission y is evaluated separately by the outcome verifier against hidden ground truth y^*. The surrounding planets denote other problem families in the Apodex Discovery registry.

repaired its approach after receiving verification feedback. What is missing is an evaluation framework whose unit of analysis is not an isolated answer, but an entire episode of tool-using, evidence-seeking, verifiable, and self-correcting problem solving.

Guided by the Apodex Discovery framework and methodology, we release **TRACES**, our reality benchmark of executable environments, episodes, and tasks with hidden verifiers, together with this paper at discovery.apodex.com.

1.1. What is a Heavy-Duty Solver?

We use the term *heavy-duty solver* for a complete system, a foundation model together with the harness, tools, memory, and control policy that drive it, charged with carrying an open-ended, real-world ambition all the way to a verifiable result (**Fig.** 1). What sets such a solver apart from a model answering a prompt is not raw reasoning power but the ability to sustain a long, stateful, and self-correcting investigation against an external world that, at many times, yields some quite surprising results. That ability can be exercised only once an ambition has been equipped with the four pieces of infrastructure named above—a problem manifest, a reality-based environment, verification, and repair. We take each in turn; the sections that follow develop the machinery previewed here in full, while a more detailed description of these important components is

provided later in this paper.

1.1.1. Problem Manifest

A heavy-duty problem begins as an ambition rather than a concrete question or task, and the first piece of infrastructure turns it into something a solver can be graded on: a *problem manifest*, the static, declarative contract that fixes everything invariant about the problem and from which concrete runnable instances are drawn. A manifest specifies:

- a **question and success criterion**—what would count as a solution. This need not be an answer known in advance, but it must be a criterion by which a proposed solution can be objectively recognized, verified, or falsified when it appears.
- a **task decomposition**—the scoped tasks the problem is broken into, each with its own input, expected output, and success criteria, so that load-bearing intermediate results—those on which the final answer depends—are checked where they arise rather than inferred from a terminal artifact. Our adeno-associated virus (AAV) capsid-design problem, for instance, decomposes into four tasks: predicting whether a variant is viable, predicting its tissue tropism, reconstructing its three-dimensional structure, and generating novel on-target sequences under a limited experimental-feedback budget.
- a **target outcome and hidden verifier**—the quantity that defines success, together with the hidden procedure and scoring rule that grade a submission against ground truth, or against a measurable proxy when the definitive outcome is delayed or unobservable.
- the **data and tools**—the inputs and the action interface the solver is granted, and nothing beyond it, held separately from the hidden ground truth on which the submission is graded.
- the **resource budgets**—the time, compute, tool-call, safety, and data-access limits under which the solver must operate.

Concrete *episodes*, which are individual runnable instances, are instantiated from the manifest, each carrying its own instance data and hidden ground truth while sharing the tools, verifier, and scoring rule the manifest declares. A solver system is therefore measured across many instances of a task rather than on a single case.

1.1.2. Reality-Based Environment

The manifest says what must be solved; the environment is where the solving happens. A reality-based environment is the stateful, interactive substrate that realizes a manifest and through which a solver acts on the problem: tools (coding, search, specialized software), data (public or proprietary), and other interactions allowed by the environment manifest. Its defining property is that it returns fresh observations in response to the solver's actions: tool outputs, execution errors, retrieved documents, simulation results, experimental measurements, or verifier feedback. An environment is therefore not a prompt. A prompt supplies static context, whereas an environment produces new information as the solver acts.

Two commitments make the environment a faithful stand-in for reality. First, *isolation*: the hidden ground truth on which a submission is graded is structurally prevented from reaching the solver, and execution is confined so that answers cannot be retrieved from outside nor results exfiltrated. Second, the unit of evaluation is the complete *solver system under test*—the foundation model together with its harness and agent loop—rather than the model in isolation, because heavy-duty work is carried out by the whole system. Solver and environment meet only through a fixed *episode interface*, which grants the inputs, tools, budget, and submission format on one side and applies the hidden verifier, hard gates, and outcome metrics on the other. Because every system and every baseline passes through the identical interface, a difference in outcome can be attributed to a specific component—the model, the harness, the agent loop, the tools, or

the handling of feedback—rather than to the environment, and the same environments serve to compare harnesses at a fixed model as readily as models at a fixed harness.

1.1.3. Verification Mechanisms

An environment is only as trustworthy as the checks applied to what happens inside it, and verification takes two complementary forms. *Outcome verification* asks what the solver produced, scoring the final submission—and, where a problem is decomposed, each load-bearing intermediate result—against success criteria that are known by the solver, but detailed scoring rubrics/formulas and data held inaccessible by the solver. For many of the problems that motivate this work the definitive check is unavailable, delayed, or only a shadow of the objective that ultimately matters, so the verifier falls back on a measurable proxy, such as a held-out experimental fitness in place of eventual *in-vivo* therapeutic benefit. Because a deliverable is often a trained model, a curated corpus, or a data pipeline rather than a direct answer, each outcome verifier also carries explicit anti-gaming gates—leakage and contamination checks that disqualify an episode outright rather than merely lowering its score—and every outcome is normalized to a common scale from 0 to 1, between a trivial baseline and the best attainable result.

Process verification, on the other hand, asks how the solver produced its result, reading the recorded trajectory rather than the submitted artifact. Process verification complements outcome verification by detecting procedural deficiencies, maintaining integrity, and may sometimes be the only feasible form of evaluation for truly open-ended questions. We instantiate it through **HDS6**, a metric that scores the trajectory along the six capabilities whose initials spell **TRACES**, defined in full below, which together separate a system able to sustain a long, self-correcting investigation from one that merely assembles a plausible account. The process verifier operates blind: it cannot consult the verified outcome, the solver's private chain-of-thought is stripped from the record before judging, and the solver's identity is withheld, so that a score reflects only the externally visible behavior of the trajectory, while a single integrity gate consults the outcome solely to detect fabrication. The two forms are complementary. Outcome scores measure success but are sparse and, for the highest-value problems, may not resolve for months or years; process scores supply the finer-grained and immediately available evidence of whether the investigation was conducted reliably, and are what allow progress to be assessed even when definitive ground truth is delayed, incomplete, or not yet available.

In this paper, we introduce Apodex Discovery, a framework for building and evaluating *discoverative AI* through heavy-duty solvers, together with its reality benchmark **TRACES**. Apodex Discovery transforms an open-ended problem into a budgeted, stateful, and partially observable environment in which the evaluated object is the complete solver system under test—including the foundation model, harness, and agent loop—operating through the tools and feedback exposed by the environment, rather than the foundation model in isolation. The framework is *interactive*: the solver acts on an environment whose state changes in response. It is *verifiable*: episode-level and task-level outcome verifiers evaluate the final submission and load-bearing intermediate outputs, respectively, against hidden ground truth or measurable proxies. It is *repairable*: structured diagnostic feedback can be returned to the solver without exposing hidden ground truth, enabling failures to be localized, corrected, and re-checked. Finally, it supports process verification: HDS6 evaluates the recorded trajectory independently of the verified outcome, allowing Apodex Discovery to distinguish a reliable investigation from a submission that succeeds without adequate evidential support.

To evaluate the process capabilities required for sustained heavy-duty work, we introduce **HDS6**, a six-dimensional process-verification metric whose dimensions spell **TRACES**: Tools, Repair, Alternatives, Coherence, Evidence, and Scope.

- **Tools**: selecting, invoking, and correctly interpreting external tools;

- **Repair**: responding to verification feedback or observed failure, correcting the underlying error, and confirming the fix;
- **Alternatives**: making competing hypotheses explicit and adjudicating among them as evidence accumulates;
- **Coherence**: maintaining state, constraints, and logical consistency over an extended horizon;
- **Evidence**: grounding claims in observations, tool outputs, data, experiments, or references; and
- **Scope**: identifying the conditions under which a conclusion holds and the boundaries beyond which it should not be applied.

We refer to the resulting six-capability evaluation as *HDS6* (Heavy-Duty Solver Six Capabilities). Outcome verification asks whether the solver ultimately succeeded; HDS6 process verification asks whether the solution was produced through a reliable, evidence-grounded, and self-correcting investigation. This distinction is particularly important for high-value real-world problems, where definitive outcome ground truth may be delayed, incomplete, or initially unavailable. Together, outcome verification and HDS6 provide complementary views of whether an AI system can not only produce an answer, but also conduct the sustained investigation required to earn one.

1.1.4. Repair Loops

The final component turns evaluation from a verdict into a loop. A repair loop returns the results of verification—computation, experiment, external evaluation, or human judgment—to the solver so that failures can be localized, corrected, and re-checked rather than merely recorded. Within an episode this already occurs: verifier feedback passed back across the episode interface is what lets a solver detect a failed check and revise the work that produced it, and the Repair capability of HDS6 measures whether it does so by targeting the actual error and confirming the fix.

Verification also supports a repair loop *across* attempts, because process verification yields a structured diagnosis rather than only a number. From the subrubric assessment, each capability that scored below its top band contributes a proposed band transition—the specific behavior a higher band would require—and from the load-bearing analysis, each faulty step contributes feedback localized to that step. These findings are synthesized into a compact *repair note* and returned to the solver for a fresh attempt, and re-scoring the repaired trajectory quantifies the effect of the guidance. The note withholds the verified outcome, the hidden ground truth, and the outcome score, so that it guides the process without disclosing the answer; repair is in this sense a fixed procedure that closes a measure–diagnose–improve loop without ever relaxing the isolation between solver and ground truth. It is this loop that makes a heavy-duty problem not merely measurable but progressively solvable.

We begin by systematically scouting real-world problems that are both consequential and suitable for rigorous AI evaluation. A two-month effort by a ten-person team of STEM Ph.D. researchers surveyed 561 industries across sixteen sectors, assembled a registry of 423 high-value problems, and selected twenty for the initial TRACES benchmark. On AAV capsid design, Apodex exceeded the task-level published state of the art across all four stages of the pipeline: viability, tropism, structure prediction, and generative design; in drug repurposing, adding the biomedical environment improved the mean normalized prediction score of GPT-5.5 and GPT-5.6-sol by 2.5 and 7.6 points, respectively, over the same closed-book backbone. Beyond measuring aggregate performance, the fixed TRACES episode interface and controlled ablations make it possible to attribute success or failure to specific components of the solver system, including the foundation model, harness, initial skill guidance, and terminal verifier-guided repair, rather than treating every failure as an undifferentiated limitation of model capability.

2. Scouting High-Value Real-World Problems

Discoverative AI is only as consequential as the problems it is directed at. Identifying suitable real-world problems for AI evaluation requires a systematic process of domain selection, problem discovery, and environment construction. Although unresolved problems are abundant across science, medicine, engineering, and industry, only a subset possesses the reasoning depth, technical structure, verification pathway, and real-world value required to evaluate heavy-duty AI systems.

> *We only care about the unknowns: they are hard, uncertain and may take years to solve. But they are highly valuable to mankind. And they are the only problems we care about.*

To identify such problems, a ten-person research team conducted a two-month survey spanning scientific research, engineering, healthcare, finance, and industrial applications (**Fig. 2 and Appendix A**). All members of the team hold Ph.D. degrees in STEM disciplines and collectively contribute expertise in scientific research, AI development, engineering, computational methods, and domain-specific problem formulation. The survey combined systematic literature review, analysis of industry reports and technical materials, participation in domain-specific meetings and workshops, discussions with researchers and practitioners, and examination of problems arising from partner programs and internal deployment efforts.

This process produced an initial collection of **423 high-value real-world problems**, each of which was reviewed for reasoning depth, technical feasibility, verifiability, availability of the necessary data and tools, expected verification latency, and potential real-world impact. The resulting benchmark is therefore not a collection of questions assembled through lightweight brainstorming, but the outcome of a structured process involving domain analysis, problem collection, expert review, verifier design, and environment construction.

2.1. Domain Selection

Apodex Discovery is designed to evaluate heavy-duty problems whose solutions cannot be obtained through simple memorization or direct retrieval. A question with a published answer may still require substantial reasoning, but performance on such a question can also reflect memorization, benchmark contamination, or sophisticated information retrieval. In contrast, when a problem has no established answer, the solver must be evaluated through the work it performs, including the evidence it collects, the tools it invokes, the hypotheses it considers, the intermediate artifacts it produces, and the final outcome it proposes.

We began the domain-selection process by constructing a master taxonomy of **561 mutually exclusive industries** spanning **sixteen sectors**. This bottom-up representation of the global industrial landscape was then cross-checked against the scientific and technological challenge areas identified by the U.S. Government Genesis Mission, providing a complementary top-down view of strategically important problems.

Each industry was assessed along three dimensions:

1. **Technical fit.** We evaluated whether the most difficult problems in the industry had an appropriate structure for a long-horizon reasoning system, considering reasoning depth, formalizability, executability, verifier fidelity, verification latency, data availability, and last-mile deployability.
2. **Value.** We evaluated whether solving the industry's hardest problems could produce either a landmark scientific or technological result or a large commercial opportunity. Rather than averaging these two forms of value, we used their maximum, such that exceptional scientific significance or exceptional commercial value was independently sufficient for a domain to qualify for further consideration.

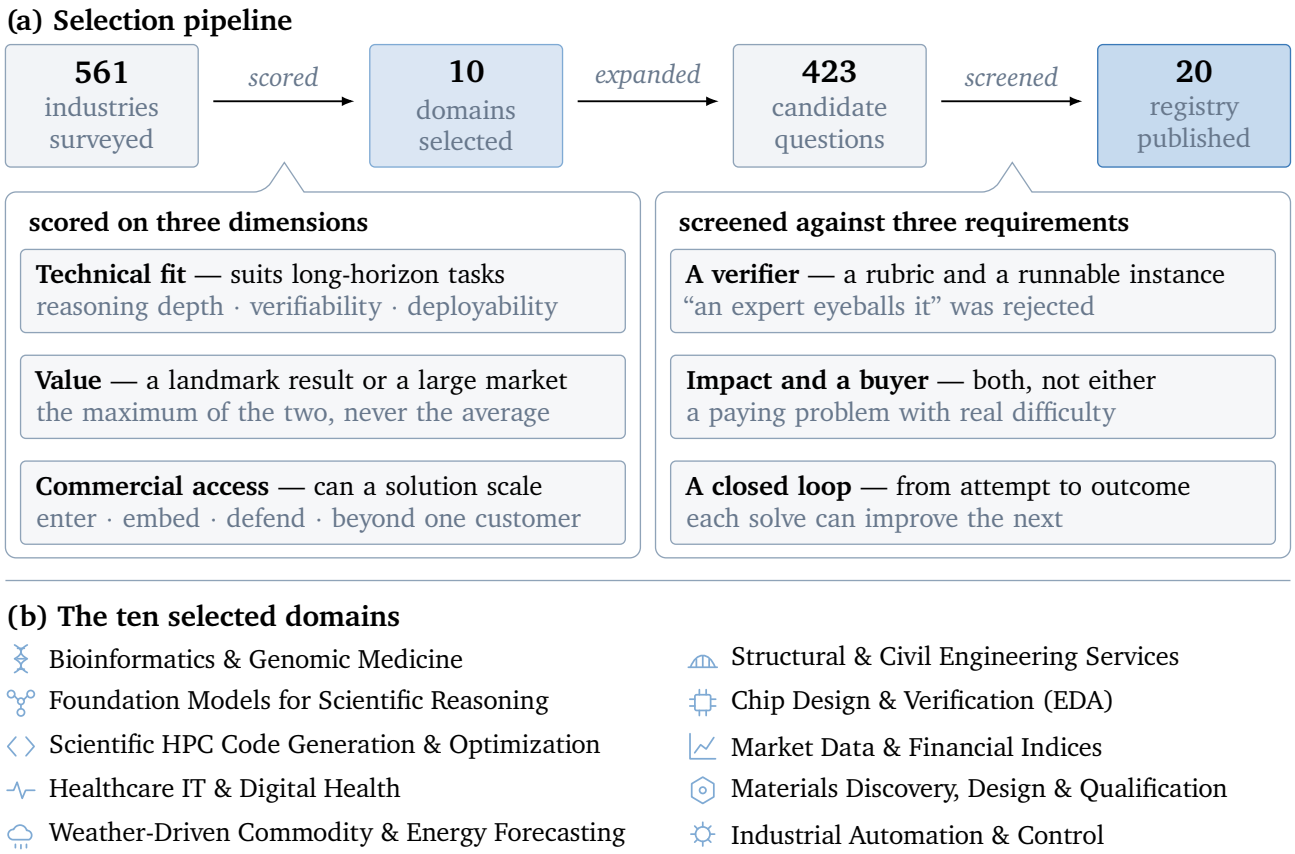

Figure 2: The selection pipeline. From 561 industries surveyed to 10 domains, expanded to 423 candidate questions and screened to the published registry of 20.

3. **Commercial access.** We evaluated whether a solution could realistically enter the domain, become embedded in an existing workflow, develop a defensible advantage, and scale beyond a single demonstration, partner, or customer.

The three dimensions were assessed independently rather than collapsed into a fully compensatory score, because, for example, strengths in market size, data availability, or commercial access should not obscure weaknesses in verifier quality or reasoning depth. Two further principles shaped the scoring. First, we judged each domain by what an advanced discoverative AI system could ultimately achieve on its hardest problems, not by how widely AI is used there today, since current adoption reveals little about that potential. Second, the depth of heavy-duty reasoning and tool use capped the overall score: a problem ranked highly only if solving it genuinely demanded such deep, multi-step work, not merely because it offered abundant data, convenient evaluation, or an established market.

Because commercial access depends heavily on institutional structure, procurement processes, data ownership, regulatory constraints, and deployment context, it was more difficult to estimate reliably than technical fit or scientific value. Quantitative scores were therefore used to organize and narrow the search rather than to determine the final selection automatically. Human reviewers then examined the highest-ranked domains and selected ten domains across eight sectors, with no more than two domains drawn from any single sector.

The resulting domains are:

- Bioinformatics, Computational Biology, and Genomic Medicine;
- AI Foundation Models for Scientific Reasoning;
- AI for Scientific HPC Code Generation and Optimization;
- Healthcare IT and Digital Health;
- Weather-Driven Commodity and Energy Market Forecasting;
- Structural and Civil Engineering Services;
- AI-Assisted Chip Design and Verification;
- Market Data and Financial Indices;
- AI-Driven Materials Discovery, Design, and Qualification; and
- Industrial Automation and Control.

2.2. Question Collection and Screening

The ten selected domains were expanded into concrete candidate problems through literature review, industry research, domain-specific meetings, expert discussions, partner programs, and internal deployment experience. The team also considered questions proposed directly by scientists, engineers, clinicians, and other domain practitioners, particularly when those questions arose from active research programs or operational workflows rather than from benchmark design alone.

For each candidate, the team documented not only the problem statement but also the data, tools, computational interfaces, verification procedures, and external dependencies required to make the problem operational. Candidates were therefore evaluated as prospective environments rather than as isolated natural-language questions, and an appealing or consequential problem description was not sufficient for inclusion unless a credible path existed from the solver's actions to an inspectable and verifiable outcome.

At minimum, each candidate problem was required to specify:

- a concrete objective and a clearly defined output artifact;
- a decomposition into substantive intermediate tasks;
- a measurable verification rubric;
- at least one runnable input–output instance;
- an identifiable source of data, evidence, or experimental feedback;
- a path from the solver's attempt to an externally verifiable outcome; and
- both a consequential impact and an identifiable user, customer, or beneficiary.

Questions whose success conditions ultimately reduced to unstructured expert judgment were rejected. Expert review may remain an important component of evaluation, particularly for frontier scientific problems, but it cannot serve as the sole basis for determining success; instead, a benchmark problem must expose sufficient observable structure for its claims, intermediate artifacts, experimental assumptions, and final outcomes to be independently inspected, challenged, and, where possible, falsified.

We also applied a strict value gate under which each problem had to demonstrate both meaningful impact and a plausible beneficiary or buyer. A problem that no identifiable party would invest resources to solve is more likely to constitute an academic exercise than a reality-facing challenge, whereas a problem with an obvious buyer but limited technical depth is more appropriately treated as a product feature than as a heavy-duty benchmark problem.

This process resulted in a registry of **423 high-value real-world problems**. From this registry, we selected **20 problems** for the initial TRACES release, spanning frontier-model development, biomedical discovery,

clinical translation, scientific engineering, and deployment-oriented intelligence.

The selected problems fall into two evaluation settings. In the retrospective setting, the benchmark creator has access to a known outcome that is deliberately withheld from the solver, and the task is constructed so that the outcome cannot be recovered through ordinary lookup but must instead be reconstructed through evidence gathering, modeling, experimentation, or tool use. In the prospective setting, represented by discovery challenges, no authoritative answer exists when the evaluation begins, and the quality of the solver's work is assessed as external evidence and real-world outcomes subsequently become available.

Although these settings differ in whether the final outcome is already known, they share the same fundamental property: the solver cannot succeed by recalling or retrieving an answer. A correct solution need not be known in advance, but it must be *objectively recognizable when it appears*, because the ability to recognize, verify, or falsify a proposed solution is what makes an open-ended real-world problem suitable for evaluation.

2.3. Reality-Facing Problems in the Initial Benchmark

The initial release of TRACES focuses on five problem families selected for their combination of substantial real-world value, long-horizon reasoning, heterogeneous tool use, meaningful intermediate verification, and external outcomes capable of ultimately confirming or falsifying the solver's conclusions. Table 1 lists the concrete tasks within each family, with brief descriptions and the number of episodes currently available.

1. **Adeno-associated virus capsid design.** This problem evaluates a four-stage pipeline for the assessment and design of adeno-associated virus capsids for gene delivery. The tasks comprise predicting whether a capsid variant remains viable, predicting its tissue or cellular tropism, reconstructing its three-dimensional structure, and generating novel, manufacturable, and on-target capsid sequences under a limited experimental-feedback budget.
   A solver must integrate sequence–function data, structural and biological constraints, predictive models, folding and analysis tools, and experimentally derived feedback. Each task is evaluated against leak-resistant held-out evidence, including extrapolation across mutational load, species, structural-deposition date, and unseen candidate sequences.
2. **Drug Repurposing and Reformulation.**
   This task evaluates the identification of new therapeutic indications and optimized formulations for existing, clinically characterized drugs, rather than the discovery of new drugs. Specifically, it focuses on repurposing drugs approved for one disease to treat different diseases. A solver must synthesize mechanistic evidence, disease biology, clinical-trial records, pharmacology, safety information, regulatory evidence, repurposing procedure and real-world outcomes to prioritize drug–disease hypotheses. The reformulation aspects include population, biomarker, drug combination, and formulation/delivery. Performance is assessed against held-out clinical, regulatory, or real-world outcomes rather than against a published answer available to the solver. The objective is not to generate a plausible therapeutic narrative, but to produce a prioritized and evidence-grounded recommendation that survives prospective or retrospective verification.
3. **Clinical-trial analysis and translation.**
   This problem covers statistical analysis, clinical reporting, safety- and efficacy-signal discovery, trial-advancement forecasting, and preclinical-to-human translation. Representative tasks include generating analysis code and tables under a prespecified statistical analysis plan, identifying emerging safety or efficacy signals, predicting whether a program will advance to the next clinical stage, and diagnosing why promising preclinical findings fail to translate to humans.
4. **Data for large language models.**

This problem addresses the acquisition, construction, filtering, and evaluation of the data underlying language-model development, with tasks including pretraining-data procurement, quality estimation, deduplication, contamination detection, benchmark construction, and the discovery of data mixtures that improve downstream capabilities under fixed compute and budget constraints.
The objective is not simply to collect more data, but to construct data pipelines whose downstream effects can be measured through controlled experiments and whose decisions can be traced to explicit evidence.

5. **Training and systems for large language models.**
This problem evaluates the discovery, diagnosis, and repair of the recipes and systems required to train and deploy language models, with tasks including synthetic-data fine-tuning, reinforcement-learning recipe discovery, optimizer and schedule selection, inference-determinism repair, distributed-training debugging, and solution-verifier construction.
Solvers must operate across codebases, logs, experiments, hardware constraints, and evaluation results, such that success requires not only proposing a training recipe or code modification but also executing it, diagnosing failures, interpreting measurements, and demonstrating that the resulting system satisfies the stated constraints.

Together, these problem families illustrate the intended scope of Apodex Discovery. They are not short-form questions expressed in domain-specific language, but extended investigations that require a solver to maintain state over a long horizon, interact with tools and data, produce inspectable intermediate artifacts, respond to verification feedback, and ultimately create work whose value can be judged outside the benchmark.

The initial executable environments represent only a small fraction of the 423-problem registry, and future releases will progressively expand into additional scientific, engineering, financial, healthcare, and industrial domains. The growing collection is intended both as an evaluation framework and as an invitation to the broader community to test models and solver systems not only on questions whose answers are already known, but also on difficult, consequential, and verifiable problems that remain to be solved.

2.4. Discovery Challenges

We define a *discovery challenge* as a reality-facing problem for which no authoritative reference answer is available when evaluation begins. Unlike a retrospective benchmark instance, in which a known outcome is withheld from the solver, a discovery challenge is evaluated prospectively as experimental measurements, external evidence, expert adjudication, or real-world outcomes become available. The solver is therefore assessed not by comparison with a pre-existing answer, but by whether it produces an auditable body of work whose assumptions, intermediate conclusions, proposed intervention, and final claims can be independently confirmed or falsified over time.

Our first discovery challenge focuses on **causal drug repurposing by functional-screen compound matching**. Given a protective-versus-detrimental gene signature from an unbiased genome-wide survival screen, in which each gene's effect on survival was established causally rather than correlatively, together with a public compound resource, the solver must produce a confidence-scored ranking of candidate compounds and, for each highly ranked compound of unestablished mechanism, a defended target and mechanism-of-action hypothesis supported by an auditable evidence chain. It must also judge whether the ranking transfers to an independent screen in a second disease; and where the evidence cannot support a committed call, that conclusion is itself scored, not forced.

Every claim is committed before any answer is revealed: each mechanistic hypothesis must trace to a citable

source verified to exist and to support it, and a claim without such grounding is scored as unsupported rather than as a hedge. A control benchmark and a withheld slice of the ranking are sealed until after a submission freeze, and the scoring metrics and baselines are fixed in advance. Because the leading candidates' mechanisms and the eventual survival outcome are not yet established, the exercise rewards a defensible causal argument over pattern-matching to the literature.

The first problem instance is a Parkinson's-disease model, where a genome-wide survival screen in disease-model neurons yields protective and detrimental gene hits and a preliminary compound ranking that already sorts sensibly, with known neuroprotectants near the top and known cytotoxins near the bottom. It addresses a structural weakness in repurposing, where reversing a disease signature that mixes pathogenic, incidental, and compensatory changes can undo a protective adaptation as readily as it corrects a harmful one. The solver must integrate the screen signature with pharmacological, chemoinformatic, and patent literature and with independent human and model-organism evidence into a calibrated ranking and a defensible mechanistic account.

Evaluation is scored by a verifier on a shared auditable execution backbone once the sealed material is revealed. Each prediction is scored against held-out ground truth rather than a correlative proxy, matching the ranking against control-set enrichment and measured survival effect, a masked-mechanism set against known targets, and the reasoning trace against a validation protocol. It is benchmarked against a conventional signature-reversal floor, a random lower bound, and the partner's preliminary ranking, run inside and outside the environment; and on a longer horizon, the top candidates are advanced to a single-cell survival assay whose outcome does not yet exist.

This challenge exemplifies the intended role of discovery challenges in TRACES: the benchmark does not assume that the correct scientific answer is already available, but it does require the solver to produce a causal argument, a confidence-scored repurposing candidate, and a falsifiable mechanistic hypothesis whose validity can be progressively tested as sealed evidence is revealed and new experiments are run.

3. Process Verification and Reality-Based Environment

In this section, we describe the verification framework and executable environment infrastructure underlying Apodex Discovery, the mechanisms through which a discoverative AI's investigation becomes verifiable rather than merely plausible. We first introduce outcome and process verification, and then formalize the environment–task–episode abstraction that supports reproducible and repairable solver evaluation. In our view, the main ingredient of our benchmark is its **verification** process: not only verifying the final output, but a complete framework that showcases verification for *partial results* on *decomposable tasks*, and a novel conceptual framework that captures agents' process integrity from six important aspects desirable for any heavy-duty solvers (HDS). After introducing our verification framework, we introduce **executable environments**, a powerful, general-purpose repository that blends *environments*, *tasks*, *episodes* and *verifiers* into a single runnable, reproducible and repairable whole.

3.1. Verification

Verification is the mechanism by which solver behavior is evaluated. We distinguish two complementary forms: *outcome verification* and *process verification*. Outcome verification evaluates what the solver produced. Process verification evaluates how the solver produced it. Both are necessary. Outcome scores measure task success, but they often provide sparse diagnostic information. Additionally, in many situations outcome verification may be difficult or time-consuming to obtain due to proprietary data issues or wet-lab verification requirements. On the other hand, process scores supply finer-grained evidence about tool use, evidence

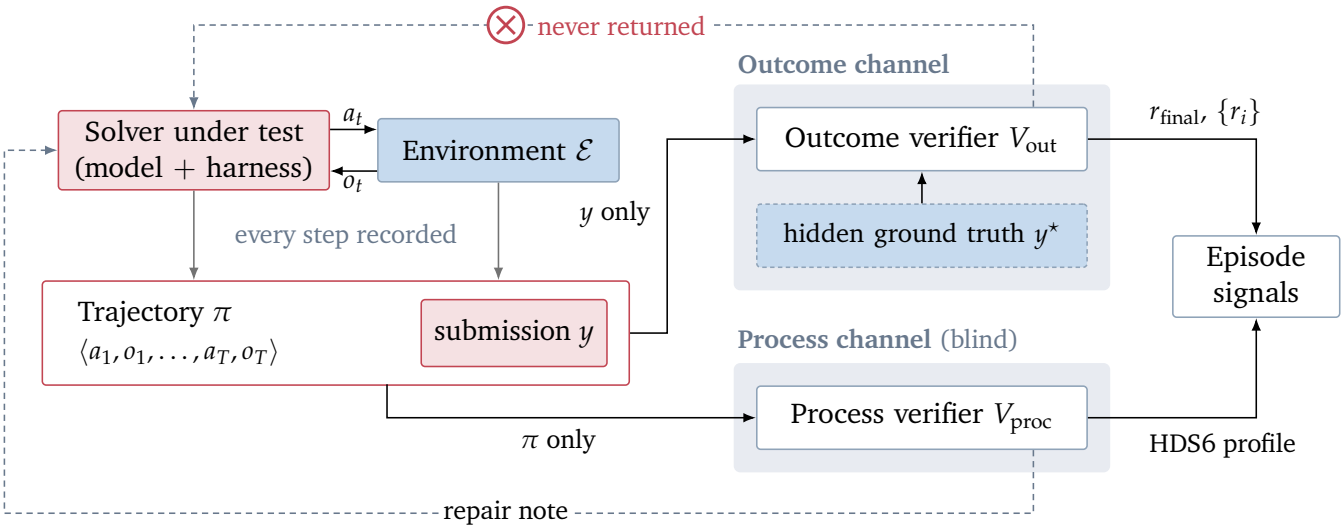

Figure 3: Verification channels. The solver and the environment interact through actions a_t and observations o_t, and their interleaved sequence is the trajectory $\pi = \langle a_1, o_1, \ldots, a_T, o_T \rangle$. The outcome verifier reads only the submitted artifact y together with hidden ground truth $y^\star$, and its result is withheld from the solver (crossed edge across the top); the blind process verifier reads the trajectory alone, without the outcome or the solver's identity, and its diagnosis does return to the solver as a repair note that still discloses neither the outcome nor $y^\star$. Both feed the episode signals: the outcome scores $\{r_i\}$, r_{final} and the HDS6 process profile.

handling, hypothesis management, repair behavior, and other capabilities that determine whether a trajectory is reliable and trustworthy.

Figure 3 summarizes the information flow across these two channels: the solver and the environment interact through actions and observations whose interleaved sequence is the trajectory, the hidden outcome verifier reads the submission against ground truth it never exposes, and the blind process verifier reads the trajectory alone.

3.1.1. Outcome Verification

Outcome verification asks what the solver ultimately produced, scoring the final submission against success criteria known to the solver using scoring procedures and ground truth that are kept hidden from the solver. The form of the check follows the task: unit tests or exact-answer keys for code and mathematics, held-out labels or benchmark accuracy for prediction problems, a simulation endpoint for a proposed policy or design, and expert adjudication where no automatic criterion suffices.

Task decomposition and proxies. An episode is seldom a single answer. Formally, an episode-level verifier scores the final submission, $V_{\text{out}}^{\text{final}}(y, y^\star) \to r_{\text{final}}$, and when a complex problem is decomposed into tasks $\tau_1, \ldots, \tau_n$ each task carries its own verifier, $V_{\text{out}}^{(i)}(y_i, y_i^\star) \to r_i$, so that important intermediate results or milestones are checked without waiting for the completion of the final artifact. In a drug-repurposing episode, for instance, the retrieved mechanistic evidence, the constructed mechanism chain, and the final promisingness estimate are each verified in turn; in a capsid-design episode for the design of adeno-associated virus (AAV), the validity of a proposed sequence under the encoding and mutation rules is confirmed before its predicted fitness is scored. Such intermediate verification gives the diagnostic resolution to localize where

Table 1: Real-world questions available in our initial benchmark.

Task	Description	Episodes
AAV capsid design		
Viability prediction	Predict whether an AAV capsid variant remains viable, meaning that it can assemble and package its genome, from its amino-acid sequence.	4
Tropism prediction	Predict the tissue- or cell-specific enrichment of AAV capsid variants, including generalization across organs, species, and human in-vitro assays.	3
Structure prediction	Predict the three-dimensional structure of an AAV capsid from sequence at the level of the surface loops, the full 60-mer assembly, and capsid–ligand complexes.	3
Generative design	Design a diverse set of novel, manufacturable, and on-target AAV capsid sequences under a limited experimental-feedback budget.	8
Drug repurposing and reformulation		
Drug–disease assessment	Given each drug–disease pair and the evidence available at prediction time, estimate its promisingness and confidence against held-out clinical and regulatory outcomes, together with a detailed repurposing and reformulation procedure.	100
Clinical trials		
Statistical analysis and reporting	Produce analysis tables, listings, and figures for a trial dataset under a pre-specified statistical analysis plan.	3
Data for large language models		
Pretraining-data procurement	Assemble a high-value pretraining corpus from candidate sources under a fixed budget while avoiding contamination.	6
Pretraining-data quality filtering	Train a scorer that predicts multiple document-quality dimensions used to filter a corpus.	2
Deduplication and decontamination	Build a pipeline that removes duplicate and evaluation-contaminating documents from a corpus.	3
Benchmark construction	Generate difficult, self-contained evaluation items that strong models cannot easily solve.	16
Training and systems for large language models		
Pretraining speedrun	Edit a training script to reach a target validation loss in the least wall-clock time.	3
Reinforcement-learning recipe discovery	Discover a training recipe that maximizes held-out gain from a limited-budget proxy.	4
Inference determinism repair	Locate and remove sources of nondeterminism in an inference stack so that outputs become reproducible.	3
SWE-agent training	Fine-tune a model to resolve real software-engineering issues.	1
Solution selection and ranking	Select or rank the correct solution among candidates without access to visible tests.	22
Solution-verifier authoring	Author a program that judges whether a model's solution to a problem is correct.	4
Operator alignment	Determine whether ported numerical kernels match their reference implementations.	33

in a long investigation a solver succeeded or failed, which a single terminal score cannot provide.

For many of the problems that motivate this work, even the final check is unavailable, delayed, or only a *shadow* of the objective that ultimately matters, and the verifier must fall back on a measurable proxy. For example, in AAV capsid engineering the aim is a vector that manufactures efficiently, evades pre-existing immunity, and transduces a target human tissue *in vivo*. Since these definitive outcomes cannot be directly observed in a dry lab, a natural verification target is a held-out experimental proxy such as packaging fitness measured by high-throughput selection [14].

Anti-gaming as part of the verifier. Because many deliverables are trained models, curated corpora, or data pipelines rather than direct answers, the cheapest way to earn a high score without genuinely solving the task is to smuggle the hidden evaluation data into the submission—for example, by training on the very examples used to grade it. Each outcome verifier for environments or tasks where such gaming is likely therefore includes an explicit leakage check. A data-curation environment measures how much of the hidden evaluation text reappears in the solver's training material and voids the episode once that overlap is too large; a deduplication environment requires the submitted pipeline to remove every copy of the evaluation data; and a data-procurement environment penalizes spam or duplicated filler padded into the assembled corpus. Each check is a hard gate, so a violation disqualifies the episode outright rather than lowering its score.

Normalization against baselines. Every outcome score is reported on a common scale from 0 to 1. Two anchors, fixed when the instance is built rather than chosen by hand, make this possible: a floor that a trivial policy would reach—a random guess or a default recipe—which maps to 0, and a ceiling representing the best result known to be attainable—an oracle or the strongest recorded solution—which maps to 1; the raw measurement is then linearly rescaled between the two. A score is thus read as a position between "no better than a trivial baseline" and "at the known frontier," rather than as a raw quantity whose magnitude is hard to compare across tasks. When several quality dimensions are combined, as in a document-scoring environment, each is rescaled against its own floor and ceiling and measured by the statistic suited to its label distribution—correlation for broad-support dimensions, mean absolute error for narrow ones, and detection scores for rare positives—so that no dimension dominates through an accident of scale.

3.1.2. Process Verification

Process verification reads the recorded trajectory rather than the submitted artifact, asking not what the solver produced but how it produced it. The trajectory is treated as an authoritative log: each step carries its action, the resource it charged, and the observation or error it returned, and rejected or malformed attempts are retained alongside successful ones.

We operationalize process verification through a single metric, HDS6, which we take to be the bedrock of heavy-duty-solver (HDS) capability, since its dimensions name the competencies that separate a system able to sustain a long, self-correcting investigation from one that merely assembles a plausible account. These six capabilities are the process competencies on which any discoverative AI must be judged. HDS6 scores a trajectory along six dimensions, denoted by the mnemonic TRACES:

- **Tools**: whether the solver selects, invokes, and interprets tools correctly. Heavy-duty work is carried out through tools, so a wrong tool, a malformed call, or a result never absorbed into or improperly interpreted by the solver's state leaves the reasoning running open-loop, disconnected from the real state of the world.

- **Repair**: whether the solver can detect and self-correct incorrect steps, or respond to verification feedback or failures in an effective manner, rather than repeating the same mistakes; such repair or self-correction capabilities are central to a discoverative AI's ability to sustain ultra long-chain reasoning and working efforts;
- **Alternatives**: whether the solver makes competing hypotheses explicit, ranks them, and updates the ranking as evidence arrives. Premature commitment to a single explanation is the dominant failure mode in open-ended discovery, and disciplined management of alternatives is how a solver avoids confident but wrong conclusions.
- **Coherence**: whether the solver maintains long-horizon state, constraints, and consistency across the trajectory. Over many steps a solver that loses track of its own commitments produces a final answer that silently contradicts its intermediate work, which should be avoided by a disocoverative AI system that can properly maintain long-term coherence over hundreds of steps or turns;
- **Evidence**: whether every claim is faithfully grounded in observations, tool outputs, references, or data. A single unsupported or fabricated claim invalidates the conclusions that depend on it, so grounding is what distinguishes a defensible result from a plausible or even hallucinated guess.
- **Scope**: whether the solver states the boundaries, uncertainty, failure modes, and applicability limits of its result. A conclusion offered without the conditions under which it holds is unsafe to act on, however correct it may be within those conditions.

Taken together, these dimensions assess whether a conclusion was *earned* by the process that produced it.

We score the trajectory against these six capabilities in two complementary ways, described below, and in both the process verifier operates blind: it cannot access the verified outcome, the solver's hidden or proprietary chain-of-thought is stripped from the record before judging, and the solver's identity is withheld, so that a score reflects only the externally visible behavior of the trajectory. There is also a single "integrity gate" that detects the most serious process deficiency such as phantom tool calls or fabricated tool call results, which once detected would invalidate the entire trajectory.

HDS6 subrubric-based scoring. The first approach scores the trajectory against a fixed checklist of these six capabilities. Each capability is separated into several subcapabilities, twenty-five in total across the six, and each subcapability carries a *subrubric*: a scoring item graded on a short ordinal band $(0/1/2)$ whose anchors are authored per task, so that the capability being measured is shared across environments while the behavior that earns each band is specific to the task at hand. Table 2 illustrates one such subrubric, claim grounding within the Evidence capability, authored for two unrelated environments: the capability, the three-band scale, and the aggregation are identical, yet the trajectory behavior that earns each band differs entirely. A separate integrity gate detects the most serious process deficiency such as phantom tool calls or fabricated tool call results, which if detected would invalidate the entire trajectory. The grading is carried out by an agentic procedure rather than a single prompt: one model proposes a band with quotes cited to specific steps, a second is required to argue the strongest case for a different band, and a third arbitrates any disagreement that remains, and every quote is re-grounded against the log so that a judge cannot support a band with text the trajectory does not contain.

Load-bearing step-based scoring. The second approach does not apply a fixed checklist but first identifies which steps of a particular trajectory actually matter. It reconstructs the run as a dependency graph and propagates influence backward from the committed answer, weighting each step by how much the final result rests on it; the steps that carry most of this weight are the *load-bearing* ones. Each load-bearing step is then re-solved independently, by a panel of models that never sees the solver's own reasoning, and the

Table 2: One subrubric, two environments. Both columns instantiate the same subrubric, *claim grounding* (a subcapability of Evidence), scored at bands 0/1/2. The capability being scored is identical; the trajectory evidence that earns each band is authored to the task.

Band	Protein engineering (model viability)	Inference nondeterminism (root-cause repair)
0	Claims model quality with no held-out evidence and no comparison to a naive baseline.	A key conclusion (root cause, or which configurations diverge or are fixed) has no supporting probe evidence, or contradicts a probe result already shown in the trace.
1	Reports a validation check but does not position it against the stated additive and distance baselines.	Has evidence but the chain is incomplete, or draws a determinacy conclusion over unswept axes (other backends, cache dtypes) from a few probed corners.
2	Grounds claims in validation performance and explicitly positions the model against the naive baselines.	Every load-bearing claim traces to a specific probe output across mode and configuration, covering confirming re-runs and residual divergences, and declares unswept axes rather than assuming coverage.

step is scored by how far it departs from that independent solution—a continuous gap rather than a discrete band. These per-step gaps, weighted by load-bearing influence and by the panel's confidence, aggregate into a process-quality score and, more usefully, into targeted feedback that localizes where a trajectory went wrong. Note that here "load-bearing steps" should not be confused with task decomposition and hidden outcome verifiers for intermediate results introduced in the previous section: load-bearing steps are organically detected and constructed by the process verifier for a single trajectory, while task decomposition and the design of verifiers for intermediate outcomes is locked in when environment is designed and is out of the control from both the system under test, or the process verifier.

Structured repair feedback. Both scoring approaches yield not only a number but a structured diagnostic that can be returned to the solver, which is what closes the measure–diagnose–improve loop essential for a discoverative AI system. From the subrubric checklist, each subcapability that scored below its top band contributes a proposed *band transition*: the band the trajectory received, paired with the specific behavior a higher band would require. From the load-bearing analysis, each faulty step contributes feedback localized to that step, whose precision tracks the re-solving panel's confidence, with a confident departure yielding a full correction stating where the step went wrong, why, what it should have been, and what to check next.

Calibration and alignment with outcome scores Albeit being primarily process centric, we also check the calibration and alignment of the HDS6 metric with outcome scores (from hidden outcome verifiers) to complement our understanding of the TRACES scores. Over the 409 judged trajectories that carry a numeric outcome we correlate each trajectory's outcome score with its HDS6 process score, pooled within each environment (Table 3) and shown trajectory by trajectory in Figure 4. All four problem families show positive correlations—Spearman ρ ranges from $+0.24$ on the AAV capsid-design pipeline to $+0.65$ on training-and-systems tasks, with comparable Pearson r, and the pooled coefficient is $+0.51$ ($r = +0.56$)—so the blind process judge and the hidden verifier largely concur on which trajectories are good even though neither can see the other. Because each task carries its own outcome verifier, these pooled figures mix per-task scales and are read as agreement summaries rather than calibrated magnitudes; taken one task at a time the agreement is often sharper, with a median ρ near 0.5 that reaches 0.8–0.9 where the task admits a clear success signal.

Table 3: Outcome–HDS6 rank correlation (Spearman ρ) and linear correlation (Pearson r) over the 409 judged trajectories that carry a numeric outcome, pooled within each environment of Table 1; Each environment pools several tasks whose outcome verifiers use different scales, so the coefficients are read as agreement summaries rather than calibrated magnitudes.

Environment	n	Spearman ρ	Pearson r
AAV capsid design	220	+0.24	+0.26
Clinical trials	13	+0.47	+0.62
Data for large language models	62	+0.46	+0.51
Training and systems for large language models	114	+0.65	+0.74
All environments (pooled)	409	+0.51	+0.56

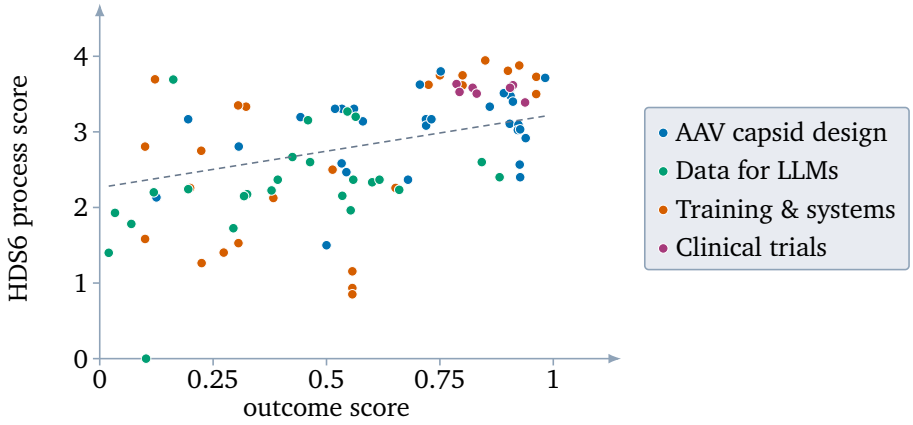

Figure 4: Each point is a judged trajectory with outcome and HDS6 scores. For clarity, fully-failed (outcome 0) and fully-solved (outcome 1) trajectories are omitted, and trajectories are subsampled to at most ten per task and coloured by problem family. The dashed line is the least-squares fit over all 283 intermediate-outcome trajectories (HDS6 $\approx 2.26 + 0.96\times$ outcome).

3.2. Environments, Tasks, and Episodes

This section describes the machinery of an *executable environment*, our foundation to a heavy-duty system for discoverative AI. There are three nested constructs—*environments*, *tasks*, and *episodes*—which, together with the verification agent teams that grade the work done inside them, make up a complete runnable, verifiable, reproducible environment. From a solver system's perspective, an environment exposes an action interface, the available tools, and the resource budgets, and it returns fresh observations as the solver acts rather than serving a fixed prompt. A task is a problem posed within an environment, representing the entire question or an important step at which final or intermediate results can be verified where they arise; an episode is a single runnable instance of that task, provisioned with its own data, budgets, and hidden ground truth, and it records the complete trajectory the solver produces. Overlaying all three are the verification agent teams of Section 3.1—a hidden outcome verifier over the submitted artifact and the blind process judges over the trajectory—which score each episode. An environment, its tasks and episodes, and these verifiers together form one runnable, reproducible, and repairable realization of an otherwise open-ended problem. The remainder of this section defines each construct in turn.

Figure 5 shows the pieces working together on one such task: two solver trajectories carried through to scoring, with their HDS6 capability profiles and one item lane resolved by the judging agents of Section 3.1.

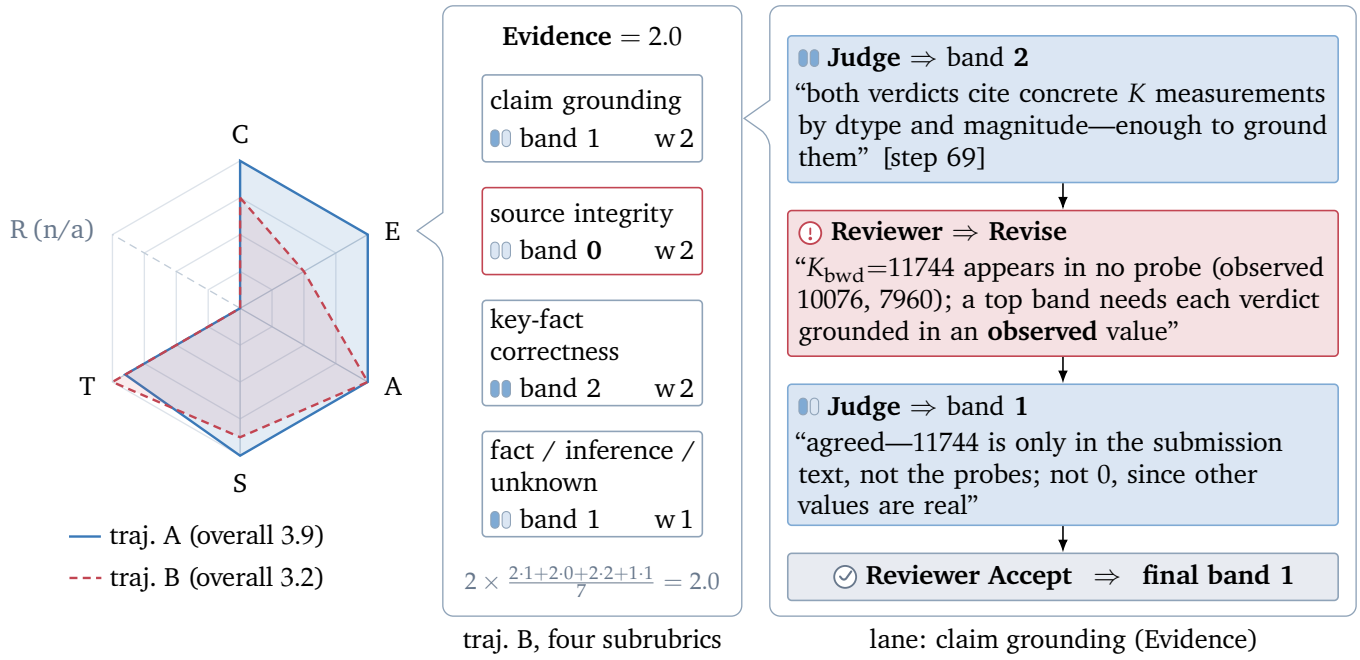

Figure 5: HDS6 on two trajectories for one operator-alignment task (does a ported numerical kernel match its reference?). Left: the six capability scores (0–4; rings mark 1–4); both trajectories pass the integrity gate, Repair is n/a for this single-submission task, and the two differ most on Evidence (4.0 versus 2.0). Middle: that Evidence score, decomposed into its four subrubrics, each scored 0–2 and weighted 2:2:2:1, whose weighted mean doubles to 2.0; the heaviest loss is source integrity at band 0. Right: the *claim grounding* lane, where the reviewer disputes the judge's top band—a cited backward figure (K=11744) appears in no probe—so the judge lowers it to 1 and the reviewer accepts.

Claim-grounding band anchors for this task. **0**: the verdict cites no concrete K value or divergence. **1**: cites a number but not enough to separate aligned from misaligned (one dtype only). **2**: each verdict carries the specific K measurements (dtype / magnitude / pass) that ground it.

Constructing an executable environment of this kind from an open-ended research problem is the labor-intensive part of this work, and we treat it as a methodological contribution in its own right: for the problems considered here, building a faithful environment with a trustworthy hidden verifier is at least as difficult as the task the solver is asked to perform.

3.3. Environments

An *environment* is the operational substrate of a benchmark or solver runtime. It may include file systems, databases, APIs, code execution sandboxes, simulators, laboratory interfaces, retrieval systems, hidden labels, and evaluator services. It also specifies resource constraints, such as time limits, compute budgets, tool-call limits, safety boundaries, and data-access rules. In practice an environment is declared by a *manifest*: a full specification of the task, the target outcome, the hidden verifier, the data provided to the solver, and the tools it is permitted to call, together with the resource budgets above. The manifest is the static contract from which the environment's episodes are instantiated with the right configurations of data, tools, verifiers, scoring rules, and other important configuration parameters.

The defining property of an environment is that it returns observations in response to solver actions. These observations may include tool outputs, execution errors, retrieved documents, simulation results,

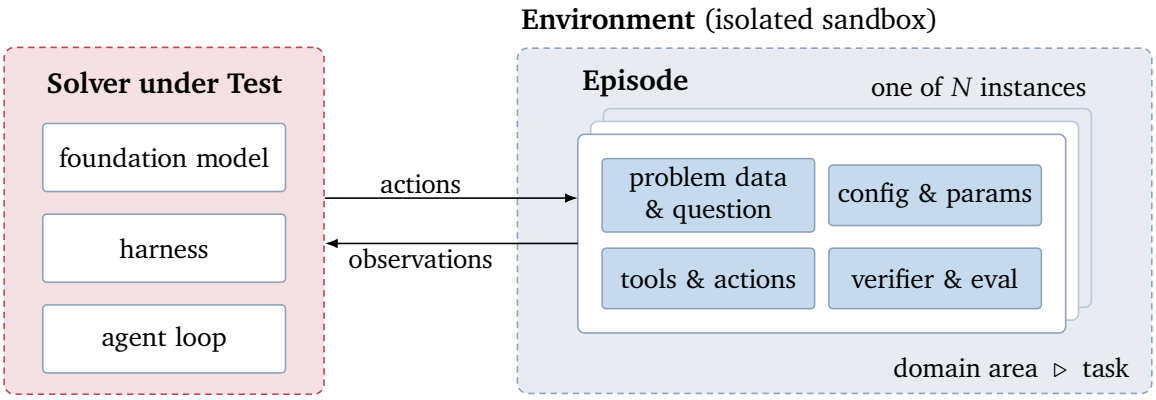

Figure 6: Environment, tasks, and episodes. An environment is an *isolated* sandbox with which a Solver under Test (SuT) interacts through actions and observations. Within it, each task or combination of tasks is realized by many episodes, its runnable instances. An episode is the concrete interface between the environment and the solver, and the ground truth stays inside the sandbox and never reaches the solver.

What each episode fixes. *Problem data & question*: the instance's inputs and the criterion for what counts as solved. *Config & parameters*: the seeds, budgets, and difficulty settings that make the instance concrete. *Tools & actions*: the action interface the solver may call, and nothing beyond it. *Verifier & eval*: the hidden grader and scoring rule applied to the final submission. The same task or combination/sequence of tasks is instantiated as many such episodes, so a system is measured across instances rather than on a single case, and each episode's full trajectory is recorded for both outcome and process verification.

experimental measurements, verifier feedback, or human-entered data. Because of this, an environment is much more than a single prompt which, even if carefully crafted, provides only static contexts, whereas an environment can produce new information as the solver acts. Figure 6 shows this structure and, in particular, what a single episode exposes to the solver.

3.3.1. Tasks

A task is a scoped work unit within an environment, defined by its input, its expected output, the tools it permits, its success criteria, and its verification procedures. Tasks are important for two reasons.

First, tasks decompose a complex real-world problem into manageable sub-problems that can be verified where they arise, rather than inferred from a single terminal result. For example, our AAV capsid-design environment decomposes the discovery pipeline into four tasks: predicting whether a capsid variant remains viable, predicting its tissue or cellular tropism, reconstructing its three-dimensional structure, and generating novel, manufacturable, and on-target capsid sequences under a limited experimental-feedback budget. Each task carries its own verifier, allowing the solver's progress and point of failure to be localized within the pipeline rather than collapsed into a single final score.

Additionally, tasks allow the definition of an *opportunity matrix* of HDS6 capabilities, which enables a more targeted evaluation of those capabilities of a discoverative AI system by declaring, for each task, which of the six are in play; Table 4 shows this variation across a range of tasks. For example, the viability-prediction task exercises *Evidence*, since a viability call must be grounded in sequence and structural features, and *Scope*, since the solver must state the regime in which its prediction holds; yet, being a single graded prediction with no feedback loop, it gives no occasion to measure *Repair*. The sequence-design task of the same environment complements it precisely there: as the solver proposes candidate sequences, checks them against the environment's viability and constraint feedback, and revises the ones that fail, it exercises *Repair* together with *Tools*. Combining the tasks of an environment therefore yields a rather complete picture of a

Table 4: Opportunity matrices for a range of benchmark tasks: different tasks activate different HDS6 capabilities. Columns are the six TRACES dimensions—**T**ools, **R**epair, **A**lternatives, **C**oherence, **E**vidence, **S**cope; ● required, ○ optional, – not applicable.

Task	T	R	A	C	E	S
Solution ranking	○	–	●	●	●	●
Capsid viability	●	–	●	●	●	●
SWE issue-fixing	●	●	○	●	●	●
Corpus deduplication	●	○	○	○	●	●
Speedrun training	●	●	○	●	○	●
Clinical trial analysis	●	●	●	●	●	●

solver system's capability as measured against the six core dimensions, even though no single task exercises or measures all of them.

3.3.2. Episodes

An episode is a concrete solver run on a specific problem instance, and it is the unit in which the solver system under test meets the environment. Within an episode the solver acts and the environment responds—returning observations, charging resources against a budget, and accepting intermediate and final submissions. Episodes are consequently the primary unit of end-to-end evaluation: task-level scores explain performance within an episode, while the episode-level outcome score grades the final result.

This meeting between solver and environment takes place only through a fixed *episode interface*, and that discipline is what gives the evaluation its meaning (Figure 7). The interface fixes, on the environment's side, the inputs, tools, budget, and submission format the solver is granted, and, on the evaluation side, the hidden verifier, the hard gates, and the outcome metrics that grade what the solver returns; the verifier feedback passed back across it is what makes an episode *repairable*. Two products cross this boundary from the solver to the evaluation layer, and the whole verification framework rests on them: the submission, which outcome verification grades against ground truth the solver never sees, and the *trajectory*, which the blind process verifier of Section 3.1.2 reads. Because the solver under test and every baseline pass through the identical interface, a difference in outcome can be attributed to a specific part of the solver system—its foundation model, harness, agent loop, tool use, or handling of verifier feedback.

The episode record. The trajectory is the core of a self-contained episode record: the ordered sequence of attempts, each with its actions and parameters, the resource charged, the remaining budget, and either an observation summary or the error raised, together with the metered model calls, any asynchronous verifier jobs, a snapshot of the final workspace, and the verification result. Such episode records can then be evaluated by the process verifier, and capability scores as well as diagnostic repair notes can be produced to evaluate and further improve the SuT's capability on the particular tasks.

4. Results on High-Value Problems

The following results are early, promising evidence for discoverative AI. Evaluating heavy-duty solvers along different tasks, we find that they surpass the published state of the art on AAV capsid design and gain measurably from a task-specific environment on drug repurposing. We also find some evidence on the effectiveness of the HDS6 metric, and its induced verify-repair loops. These are encouraging first steps toward the investigations from which genuine discoveries emerge.

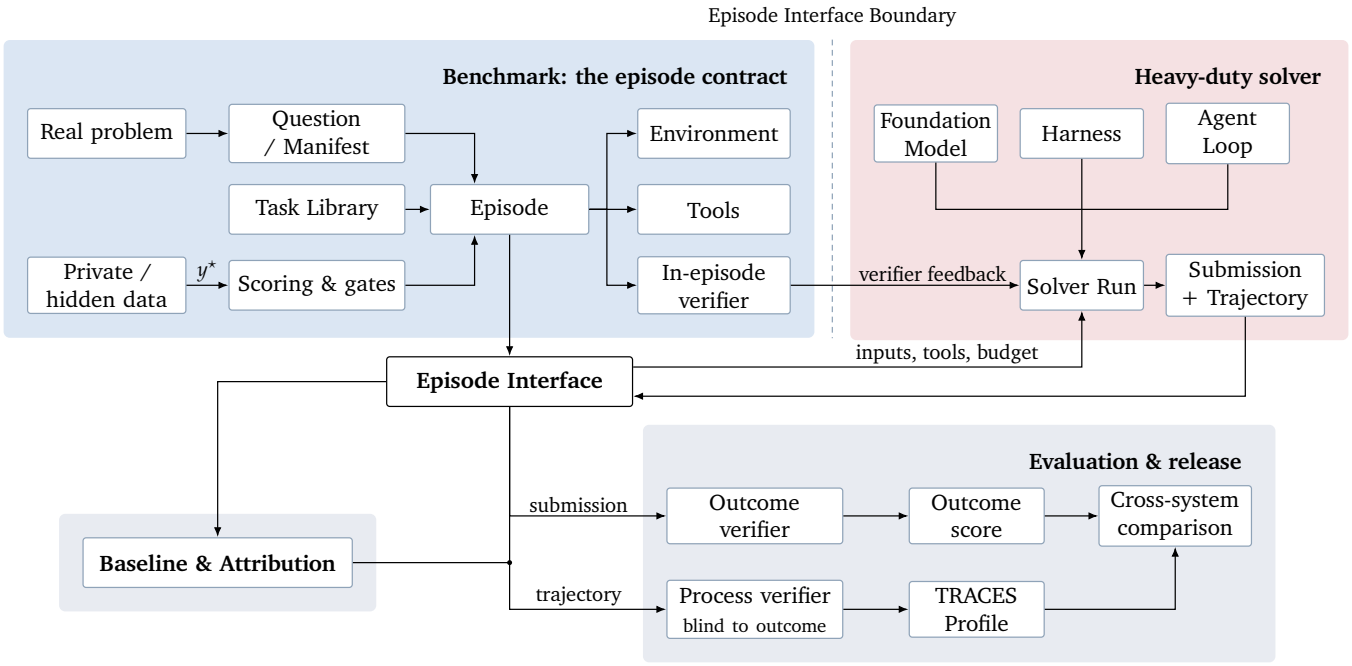

Figure 7: The framework across the episode-interface boundary. The benchmark side defines the episode contract; the heavy-duty solver consumes it and returns a submission with its trajectory; the evaluation layer is outcome-first, with the TRACES/HDS6 process profile as a diagnostic. The solver under test and all baselines pass through the same interface, which is what makes attribution to a component of the system, rather than to the environment, meaningful.

4.1. Adeno-associated virus capsid design

4.1.1. Task Formulation

Adeno-associated virus (AAV) is widely used as a leading gene therapy vector for treating a range of genetic diseases, including inherited retinal disorders, hemophilia, muscular dystrophy, and neurological conditions. Its protein shell, the capsid, determines whether the vector can be manufactured and whether it can efficiently reach target cells. Capsid engineering is challenging because mutations that improve delivery or reduce immune recognition may also prevent the virus from assembling or packaging DNA. Our benchmark captures four key stages of this process: Task 1 predicts whether a capsid variant remains viable [14, 15]; Task 2 predicts which tissues a variant targets [16]; Task 3 predicts the capsid's three-dimensional structure from its sequence; and Task 4 designs a small, diverse set of novel, on-target sequences for laboratory testing [16, 17].

Together, these tasks form a realistic discovery pipeline, in which a candidate must be manufacturable, reach the right target, and be novel enough to improve on existing vectors, all under a limited experimental budget. The benchmark therefore evaluates not only whether an AI model can predict protein properties, but whether it can support the sequence of biological decisions required to turn computational designs into experimentally testable AAV candidates.

Rigorous control of data leakage is essential to the validity of any benchmark that aims to measure genuine biological reasoning rather than memorization of the underlying corpora. We therefore built every task on real screening or structural data and enforced out-of-distribution splits across mutational load, species, and structural deposition dates. Ground truth is further protected because it is exposed only through a metered

scoring action, rather than as labels the solver can read directly. Full task definitions, metrics, splits, and per-instance result tables are given in Appendix B.

4.1.2. A domain-specific environment improves outcomes

The leading AI model operating within our domain-specific environment consistently outperformed the same model using the generic Claude Code harness, demonstrating the value of providing models with task-specific data, tools, biological constraints, and executable feedback. Averaged across tasks 1-3, claude-opus-4-8 achieved a score of 0.741 in our environment, substantially outperforming its score of 0.716 with the Claude Code harness (Table 5). The gap was widest on structure prediction (0.657 versus 0.595), the stage that most depends on orchestrating folding engines and validating intermediate results rather than returning a single prediction. The same advantage held on Task 4 (Table 19), indicating that the benefit of a domain-specific environment is not confined to any single stage but recurs throughout the capsid-engineering workflow.

4.1.3. Surpassing the published state of the art

Building on this domain-specific environment, apodex-1.1 surpassed the previous published state of the art at every stage, from predicting whether a variant survives to designing new candidates from scratch (Table 16-19). The viability task first asks whether an engineered capsid can assemble and package its genome, the first and rate-limiting filter of any campaign. On it, apodex-1.1 reached an out-of-distribution AUROC of 0.904, above the CAP-PLM capsid language model that marks the published state of the art (0.878) [14, 18–20]. This result establishes that apodex-1.1 can build AI models that reliably identify which candidates are worth carrying forward before any downstream analysis begins.

Beyond viability, two further dimensions determine whether a candidate is useful, the tissues it reaches and its three-dimensional structure. The tropism task predicts which tissues a variant reaches and whether this targeting pattern holds across species. Here, apodex-1.1 scored 0.635, again outperforming the published state of the art (Fit4Function, 0.622) [16], demonstrating that apodex-1.1 has developed models that can rank variants by tissue enrichment more accurately than this established expert method. The structure task asks the solver to build the three-dimensional capsid at three levels, the surface loops of a single subunit, the full 60-mer shell, and the capsid in complex with an antibody or receptor, which together form the basis for epitope mapping and receptor-guided design. Averaged over the three levels, apodex-1.1 scored 0.649, above the published state of the art of 0.605, AlphaFold 3 [21] with template-based symmetry expansion.

The design task, by contrast, requires a fundamentally different, generative capability. The solver must search a vast sequence space, balance viability against specificity and novelty, and prioritize candidates under a limited experimental budget. On this task, apodex-1.1 scored 0.180, surpassing the specialist generative methods AAVGen (0.109), AAVDiff (0.116), and ALICE (0.110) [19, 22, 23]. This result shows that apodex-1.1 can generate viable, novel, and on-target candidates under a constrained experimental budget. Together, these results suggest that frontier models equipped with an appropriate scientific environment can not only automate existing analytical workflows, but also make higher-quality biological design decisions than established expert approaches.

4.1.4. Results based on process verification

The HDS6 process scores exhibit the same overall trend as the outcome-based results, since systems that achieved stronger final biological performance also received higher scores for their scientific problem-solving process. Within the domain-specific environment, stronger-performing models generally showed more coherent hypothesis formation, more effective use of data and tools, better handling of biological constraints, and stronger self-correction. The same pattern holds at both the task and instance level. For instance, on the

Table 5: Model comparison across the four AAV tasks, ordered by mean result score over Tasks 1–3. For each task, we report the result score and the mean HDS6 process score across instances. All methods are evaluated in our domain-specific environment, except CC Opus-4.8, which uses the Claude Code harness.

	T1 Viability		T2 Tropism		T3 Structure		T4 Design		
solver	score	HDS6	score	HDS6	score	HDS6	score	HDS6	**mean T1–T3**
kimi-k3	**0.936**	3.49	0.649	3.61	**0.679**	**3.68**	**0.194**	**3.59**	**0.755**
claude-opus-5	0.936	3.46	**0.653**	3.55	0.668	3.61	†	-	0.752
glm-5.2	0.932	3.27	0.644	**3.76**	0.665	3.39	0.191	3.32	0.747
claude-opus-4-8	0.931	**3.65**	0.635	3.30	0.657	3.36	†	-	0.741
apodex-1.1	0.904	3.36	0.635	3.37	0.649	3.52	0.180	3.23	0.729
CC opus-4-8 (w/o env.)*	0.921	-	0.633	-	0.595	-	†	-	0.716
deepseek-v4-pro	0.918	3.37	0.613	2.97	0.610	2.19	0.129	2.64	0.714
claude-sonnet-4-6	0.922	3.45	0.628	2.77	0.575	2.73	0.094	2.97	0.709
gpt-5.5	0.909	2.99	0.600	3.57	0.544	1.22	0.142	2.24	0.684
qwen3.5-397b	0.896	3.27	0.605	2.93	0.543	1.97	0.085	2.21	0.682
apodex-1.0	0.884	3.23	0.518	2.90	0.526	1.84	0.072	2.34	0.643
published SOTA**	0.878	-	0.622	-	0.605	-	0.116	-	0.702

* the Claude Code agent running on claude-opus-4-8 without using our domain-specific environment.
** published SOTA is, per task, the strongest reference method on that task, scored by its mean over instances.
† Opus-series models (claude-opus-5, claude-opus-4-8) declined to produce outputs on Task 4, consistently triggering the models' biosafety refusal behavior.

structure prediction task (Task 3), gpt-5.5 scored only 0.544, 20% lower than kimi-k3, and its process scores fell to 1.38 on the single-subunit loop instance and 0.94 on the full-shell assembly instance (Table 18), due to the fact that gpt-5.5 consistently skips the self-validation step the task requires and submits its first-pass answer instead. Kimi-k3, the strongest solver on structure, earned some of the highest process scores on the same two instances, reaching 3.56 and 3.60, respectively. This consistency suggests that HDS6 can meaningfully characterize the quality of scientific problem solving, rather than merely evaluating the fluency or presentation of the final response.

4.1.5. Datasets and evaluation

Each task is built from published screens or structures, and is split into instances that each test a different perspective; the exact metrics are defined in Appendix B. The viability task draws on the deep-mutagenesis screen of Bryant et al. [14] and the single-mutation scan of Ogden et al. [15]. It has two instances. The multi-mutation instance tests whether a model trained on lightly mutated variants can predict the viability of heavily mutated ones; the single-mutation instance tests whether it generalizes to single substitutions at positions across the capsid sequence that it has not seen in training. Both are scored by how well the model separates viable from non-viable variants that carry the same number of mutations.

The tropism task uses the Fit4Function screens [16], where a seven-residue peptide is inserted into a fixed surface loop of AAV9 and its enrichment in each target tissue is measured. It has three instances: multi-organ tropism in mouse, cross-species transfer to macaque, and human in-vitro assays. Predictions are scored by their correlation with the measured enrichment.

The structure task gives only the amino-acid sequence and asks the solver to build the three-dimensional capsid structure. It has three instances for different structural levels: the variable surface loops of a single subunit, the full 60-mer icosahedral shell, and the complex with an antibody or cell-surface receptor. All target structures were deposited after 30 September 2021, later than the folding engines' training cut-off, so they could not have been memorized during pretraining. Each instance is scored against its experimental reference structure.

The design task asks the solver to generate novel peptides for a target, under a limited query budget. Its seed data and scoring oracles are built from the Fit4Function screens [16], and for the receptor instances from LY6A and LY6C1 binding measurements [17]. A generated sequence counts only if it is novel, manufacturable, and on target, meaning it is either selectively enriched in the intended organ or selectively binds the intended receptor.

4.2. Drug Repurposing and Reformulation

4.2.1. Task Formulation

Drug repurposing and reformulation (DRR) does not aim to discover new drugs. Instead, it seeks to identify new therapeutic indications and optimized procedure design for existing approved or clinically characterized drugs, enabling their application to additional diseases and patient populations. In contrast to de novo molecule discovery, DRR can reuse prior pharmacology, safety, manufacturing, and clinical evidence. The task includes not only indication selection, but also biomarker-guided patient stratification, formulation and delivery, combination therapy, and protocol design.

DRR decisions require the integration of heterogeneous and sometimes conflicting evidence. Mechanistic plausibility alone is insufficient because biological hypotheses may not translate into clinical benefit, trial reports may be incomplete, and outcomes depend on dose, safety, population selection, delivery, combination strategy, and study design. We therefore formulate DRR as a prospective pointwise forecasting task. Given a disease and a set of candidate drugs, the system evaluates each drug–disease pair using only information available before a fixed temporal cutoff. For each pair, it outputs a promisingness score $\hat{y} \in [0,1]$, a confidence estimate $c \in [0,1]$, and an evidence-grounded rationale. The promisingness score represents the predicted position of the pair on the clinical-development and approval spectrum.

Ground-truth labels are constructed from clinical and regulatory outcomes observed after the information cutoff. The primary endpoint evaluates $\hat{y}$; confidence and rationale support complementary calibration and trajectory analyses. The benchmark contains a 100-pair public development set and a 200-pair hidden test set. Drug and disease identities are normalized before splitting, and outcome-bearing fields from trials used to construct labels are withheld from model inputs. Temporal filtering and pair-level decontamination prevent the same outcome evidence from appearing in both the model context and the evaluation label.

4.2.2. Scoring Methods

Let the evaluation set be $\mathcal{D} = \{(x_i, y_i, \tau_i)\}_{i=1}^{N}$, where x_i is the drug–disease pair, $y_i \in [0,1]$ is its observed value on the clinical-outcome scale, and $\tau_i \in \{\text{exact}, \text{lower bound}\}$ specifies how that value should be interpreted. Exact labels represent terminal outcomes, whereas lower-bound labels record a stage already attained without asserting that the program will progress no further. A failed program receives the value of the highest phase it had previously passed. The model outputs $\hat{y}_i \in [0,1]$ without observing either y_i or τ_i. Evaluation maps the observed outcome to a fixed nonuniform tier scale, applies an absolute-error loss that distinguishes exact outcomes from lower bounds, and normalizes the mean score between a knowledge-free random predictor and a perfect predictor.

Clinical outcome	Tier value	Historical count	Historical fraction
Preclinical only or failed in Phase I	0.00	662	6.3%
Passed Phase I	0.15	2,048	19.6%
Passed Phase II	0.45	2,178	20.9%
Passed Phase III	0.65	1,641	15.7%
Approved for the target indication	1.00	3,898	37.4%

Table 6: Tier-scale values and their support among the 10,427 concluded programs in the historical corpus.

Tier scale. We represent clinical outcomes using five ordered tiers: preclinical only or failed in Phase I, passed Phase I, passed Phase II, passed Phase III, and approved for the target indication. Rather than imposing equal spacing between stages, we fix a nonuniform scale informed by the empirical distribution of concluded programs in a historical corpus. The scale is specified before evaluation and is not refit for individual systems or runs. Table 6 reports both the tier values and their historical support.

Forward-looking absolute-error score. Lower-bound labels are used for successful but not-yet-approved programs and for ongoing programs. The per-item loss is

$$\ell_i = \begin{cases} |\hat{y}_i - y_i|, & \tau_i = \text{exact}, \\ \max(y_i - \hat{y}_i,\, 0), & \tau_i = \text{lower bound}. \end{cases} \tag{1}$$

Thus, exact outcomes are penalized symmetrically in proportion to the absolute prediction error, whereas a lower-bound outcome is penalized only when the prediction falls below the stage already attained. In particular, predictions above a lower bound incur no loss.

The corresponding per-item quality is

$$q_i = \max(0,\, 1 - \ell_i). \tag{2}$$

Because $y_i, \hat{y}_i \in [0,1]$, the loss lies in $[0,1]$, so the clipping operation in Eq. (2) is inactive and $q_i = 1 - \ell_i$. The raw continuous score is the plain mean over all evaluation items, including both exact and lower-bound labels:

$$S = \frac{1}{N}\sum_{i=1}^{N} q_i = 1 - \frac{1}{N}\sum_{i=1}^{N} \ell_i. \tag{3}$$

Normalized continuous score. The random-guess anchor is defined using a knowledge-free predictor R. For each item in a predefined benchmark reference set, R draws a prediction independently from the empirical training-label distribution. The random-anchor score S_R is its expected quality under the same exact or lower-bound loss branch used for that reference item. This construction accounts for both the frequency of the prediction tiers and the mixture of exact and lower-bound outcomes. The anchor S_R is computed once using R on the development set's labels, and held fixed for all systems. For the benchmark reported here, $S_R = 0.86925$. The oracle anchor is $S_O = 1$, corresponding to zero loss on every item.

The raw score is rescaled through the linear anchor map

$$S_{\text{norm}} = \frac{S - S_R}{S_O - S_R} \times 100. \tag{4}$$

Accordingly, $S_{\text{norm}} = 0$ corresponds to the expected performance of the random predictor and $S_{\text{norm}} = 100$ to perfect prediction; scores below 0 indicate performance worse than the random anchor. We call S_{norm} the *normalized continuous score* and use it as the primary outcome metric. Unlike a categorical metric, it retains the magnitude of prediction errors on the full $[0, 1]$ scale.

Categorical accuracy. For a complementary categorical measure, each prediction is assigned to its nearest tier in $\mathcal{A} = \{0.00, 0.15, 0.45, 0.65, 1.00\}$. An exact midpoint is assigned to the lower tier. Let $\tilde{y}_i \in \mathcal{A}$ denote the tier assigned to $\hat{y}_i$. Per-item categorical correctness is

$$m_i = \begin{cases} \mathbf{1}[\tilde{y}_i = y_i], & \tau_i = \text{exact}, \\ \mathbf{1}[\tilde{y}_i \geq y_i], & \tau_i = \text{lower bound}. \end{cases} \tag{5}$$

Thus, terminal outcomes require an exact tier match, whereas ongoing or successful unapproved programs are counted as correct when the predicted tier is at least the stage already attained. Missing or invalid predictions count as incorrect. We report

$$A_{\text{cat}} = \frac{100}{N} \sum_{i=1}^{N} m_i \tag{6}$$

as *categorical accuracy*. This metric is easy to interpret but discards within-category distances; the normalized continuous score remains the more fine-grained primary measure.

4.2.3. Public-100 Benchmark Results

We use two matched evaluation conditions. In the *no env* condition, the backbone receives the task input and produces a single response without external retrieval, executable tools, or network access. In the *with env* condition, the same backbone can iteratively use an executable DRR environment that provides structured biomedical knowledge, literature-derived evidence, and scientific tools. The backbone model, benchmark items, output schema, and scoring procedure are otherwise held fixed, and neither condition has access to held-out outcomes. Thus, "environment" denotes the external evidence and tool layer rather than a separate prediction model.

We evaluate GPT-5.5 and GPT-5.6-sol with medium reasoning effort as the backbones. As shown in Table 7, the *with env* condition improves every reported metric for both backbones. Agreement between the continuous and categorical metrics indicates that the result is not an artifact of normalization. Because the comparison uses three runs on the public development set, these gains should be interpreted descriptively rather than as definitive estimates of test-set performance.

4.2.4. HDS6 Trajectory Analysis

We use HDS6 to characterize the quality of the scientific reasoning process in addition to final-answer accuracy. HDS6 rates six dimensions on a 0–4 scale: long-horizon state coherence (C), evidence fidelity (E), hypothesis management (A), boundary and failure reasoning (S), tool use and execution-state management (T), and self-correction under verification (R). Scores of 0, 1, 2, 3, and 4 correspond to absent, flawed, insufficient, good, and excellent behavior, respectively. An automated process evaluator scores the recorded reasoning and tool interactions with the held-out outcome hidden. Capability scores retain the rubric's critical-failure caps. A separate integrity gate excludes trajectories that are too incomplete or malformed for reliable process scoring, and a capability marked not rated (NR) is omitted rather than assigned a score of zero. Table 8 reports capability means, dispersion, and paired condition gains.

Table 7: Outcome-prediction performance on the 100-pair public development set. The normalized continuous score is the random-anchored 0–100 metric in Eq. (4); the raw continuous score is the corresponding mean per-item quality in Eq. (3). Categorical accuracy is defined in Eq. (6). Values are mean ± sample standard deviation over three independent runs. Absolute gain is the with-env mean minus the no-env mean, computed from unrounded values; gains are reported as normalized-score points, raw-score units, and categorical-accuracy percentage points (pp), respectively.

Backbone	Condition	Normalized continuous score	Raw continuous score	Categorical accuracy
GPT-5.5	No env	53.78 ± 3.22	0.93957 ± 0.00421	$(81.00 \pm 1.73)\%$
	With Apodex-env	56.30 ± 1.56	0.94287 ± 0.00204	$(83.67 \pm 1.15)\%$
	Absolute gain	$+2.52$	$+0.00330$	$+2.67$ pp
GPT-5.6-sol	No env	54.21 ± 2.20	0.94013 ± 0.00287	$(80.67 \pm 1.15)\%$
	With Apodex-env	61.81 ± 3.92	0.95007 ± 0.00512	$(85.00 \pm 1.73)\%$
	Absolute gain	$+7.60$	$+0.00993$	$+4.33$ pp

Table 8: HDS6 capability scores on matched drug–disease cases. Condition rows report mean ± population standard deviation on the 0–4 scale; NR and missing capability ratings are omitted. Absolute gain is the mean paired within-case difference (with env minus no env) among cases rated in both conditions. Repair gains are unavailable because repair was not observable in the no-env trajectories.

Backbone	Condition	*C*	*E*	*A*	*S*	*T*	*R*
GPT-5.5	No env	1.62 ± 0.21	1.52 ± 0.32	2.35 ± 0.65	2.72 ± 0.88	1.00 ± 0.00	—
	With Apodex-env	2.76 ± 0.72	3.10 ± 0.62	2.69 ± 0.64	2.96 ± 0.65	3.64 ± 0.69	3.45 ± 0.79
	Absolute gain	$+1.13$	$+1.58$	$+0.34$	$+0.24$	$+2.64$	—
GPT-5.6-sol	No env	1.67 ± 0.18	1.42 ± 0.32	2.48 ± 0.72	2.85 ± 0.79	0.99 ± 0.15	—
	With Apodex-env	2.81 ± 0.82	3.10 ± 0.65	2.72 ± 0.62	3.09 ± 0.62	3.56 ± 0.79	3.83 ± 0.35
	Absolute gain	$+1.14$	$+1.68$	$+0.25$	$+0.24$	$+2.57$	—

Across backbones, the environment most strongly improves tool use, evidence fidelity, and long-horizon coherence, while gains in hypothesis management and scope control are smaller. This pattern suggests that the environment primarily helps systems ground claims, preserve state, and execute evidence-seeking workflows. The low no-env tool scores reflect the deliberate absence of executable tools, not failed attempts to use them.

GPT-5.5 and GPT-5.6-sol exhibit similar with-env profiles, indicating that the process-level benefits are stable across the two GPT backbones. Repair scores are conditional on an observable correction opportunity and are unavailable for the no-env trajectories, so they should not be treated as directly comparable coverage-matched estimates. The HDS6 findings are descriptive and remain subject to the survivor-selection limitation above.

4.2.5. Repurposing and Reformulation Procedure Design

We separately evaluate whether a system can propose a clinically actionable development procedure for a drug–disease pair. The system predicts four structured fields: target population, biomarker strategy, combination therapy, and formulation or delivery. Reference values are derived from registered trial protocols. In the language-model-only condition, these fields and their source records are withheld during prediction. An automated semantic evaluator compares each predicted field with its reference on a 1–5 scale, where 1 denotes an irrelevant or contradictory proposal, 3 denotes a directionally correct but incomplete proposal, and 5 denotes agreement on the key clinical content. Field means are computed over examples with an

Table 9: Drug repurposing and reformulation procedure-design results. Scores are mean semantic-evaluator ratings on a 1–5 scale. LM + context differs from LM only for population, the field for which structured trial context is available.

Field	LM only	LM + context	Δ
Population	2.89	3.78	+0.89
Biomarker	4.00	4.00	0.00
Drug combination	3.75	3.75	0.00
Formulation/delivery	4.22	4.22	0.00
Macro-average	3.72	3.94	+0.22

available reference value, and the macro-average gives equal weight to the four field means.

Table 9 compares a language-model-only condition (LM only) with a knowledge-augmented condition (LM + context). The augmented condition revises the initial model prediction using structured eligibility context retrieved from registered clinical-trial records. This context is available at sufficient coverage only for population specification; the biomarker, combination, and formulation/delivery predictions are therefore identical across conditions. Because the retrieved eligibility record is also the source from which the population reference is derived, the population comparison measures evidence retrieval and synthesis rather than fully prospective procedure generation. Population improves by 0.89 points, increasing the macro-average from 3.72 to 3.94. Together, the outcome, trajectory, and procedure-design evaluations assess three complementary capabilities: selecting promising therapeutic candidates, producing evidence-grounded reasoning, and specifying clinically actionable development strategies.

4.3. Large language model problems and ablations

Previous sections focused on problems whose verification relies on slow experimental or clinical evidence. This section turns to a different class of heavy-duty tasks: the engineering work underpinning foundation models such as pretraining data curation, filtering, deduplication, inference-serving correctness, reward design, and training-recipe optimization. These tasks are routinely delegated to research engineers because they demand sustained, tool-mediated, self-correcting investigation, not a single prompt or inference call, and they already account for much of the labor a heavy-duty solver would need to absorb in an AI research organization. This environment family also offers a practical advantage for ablations: its outcome verifiers (*e.g.,* held-out loss, throughput gates, decontamination audits, and bit-exactness suites) are recomputable on demand, enabling denser and more controlled ablations than biomedical environments allow. We use this LLM environment family to pose two nested attribution questions.

We design two hierarchical ablation layers: (i) *solver-configuration* ablation, which fixes episode interfaces while varying back-end harnesses and foundation models; (ii) *episode-mechanism* ablation, which locks environments and solver setups to toggle individual internal mechanisms on or off. We evaluate two episode-level interventions: initial skill guidance and post-submission verifier-driven repair. This decomposition quantifies performance variance stemming from the solver framework itself versus in-episode or out-episode, process-centric guidance feedback.

4.3.1. Solver-configuration ablations

The current result matrix spans 11 evaluated LLM environments, six harnesses, and up to six foundation models on the multi-model harnesses. Two harnesses are model-restricted by construction rather than by

Table 10: Aggregate system-level performance across all tested harness–model configurations on the seven pre-specified diagnostic environments. Mean score: average performance of each model across all harnesses it was paired with. Best cell: highest score achieved by any harness–model configuration for that model. Failed cells: total number of outright failures per model (score of zero or run with no valid output); note that GPT-5.6-sol and Opus-4.8 are evaluated on 35 cells while the remaining four models are evaluated on 28, so these counts are not directly comparable in rate terms across the two groups. Harness coverage is asymmetric by design: Opus-4.8 and GPT-5.6-sol include dedicated packaged harnesses (Claude Code, Codex CLI) in addition to the four model-agnostic harnesses, while other models are evaluated only on the four universal harnesses. This table summarizes the overall achievable end-to-end performance per model.

Model	GPT-5.6-sol	Opus-4.8	GLM-5.2	Kimi-K3	DeepSeek-v4-pro	Apodex-1.0
Mean score	0.588	0.575	0.516	0.553	0.478	0.459
Best cell	0.950	0.850	0.850	0.900	0.776	0.800
Failed cells	0	0	1	1	3	2

choice: Claude Code is a closed-source packaged CLI that exposes no source to instrument, so we drive it only with Opus, and Codex CLI is similarly run only with GPT, since it is tightly coupled to the OpenAI Responses API and is operated as a packaged binary rather than a modified harness. The remaining harnesses, DeerFlow, OpenHands [24], A-Evolve [25], and ApodexHarness (our own harness), are model-agnostic by design and are swept across all six models where the environment permits.

This asymmetry means the matrix cannot be summarized by a single harness-level average without first controlling for which models each harness was paired with, since an uncontrolled average would cause the performance of Claude Code and Codex CLI to be overestimated for never having been tested against weaker models. To enable reliable controlled comparisons and fine-grained ablation analyses, we defined several quantifiable inclusion criteria prior to experiment execution to select a pre-specified diagnostic subset of environments: (i) full coverage across all four model-agnostic harnesses; (ii) no severe floor-level scores; (iii) sufficient run-to-run reliability. Seven of the eleven LLM environments meet all three criteria and are used for the aggregate analyses below: `LLM-pretrain-data-probe`, `LLM-rollout-verify`, `operator-align`, `rl-recipe`, `posttrain-data-dedup`, `pretrain-data-filter`, and `solution-verifier`. All 11 environments are reported in full in Appendix C.

Model performance: aggregate scores and caveats. Even the strongest model–harness combinations remain far from solving this environment family. Across all tested configurations, the highest aggregate mean per model is 0.588, as shown in Table 10. This gap is itself a meaningful benchmark for progress, indicating that substantial headroom remains for heavy-duty solvers relative to task-specific score ceilings.

As noted in the table caption, these aggregate scores reflect end-to-end system-level performance across all tested harness pairings, including dedicated packaged solver tools. Harness coverage is intentionally asymmetric, so this table characterizes practical real-world deployment performance per model rather than isolating pure foundation model capability. To isolate pure foundation model effects with equal harness coverage, we also compute balanced means using only the four model-agnostic harnesses (a fully matched 4×6 matrix across all seven environments). Under this balanced setting, GPT-5.6-sol ranks first (0.595), followed by Opus-4.8 (0.573) and Kimi-K3 (0.553), then GLM-5.2 (0.516), DeepSeek-v4-pro (0.478), and Apodex-1.0 (0.459). Adding each model's dedicated packaged harness in the full end-to-end table changes

Table 11: Harness performance comparison on the seven pre-specified diagnostic environments, presented for two representative foundation models (Opus-4.8 and GPT-5.6-sol). Claude Code and Codex CLI are dedicated packaged harnesses tightly coupled to a single model, and are evaluated only on their respective native model. DeerFlow, OpenHands, A-Evolve, and ApodexHarness are model-agnostic harnesses designed to work across all foundation models; the two representative configurations shown here illustrate how relative harness performance varies with model choice. "Best cell" is the highest score the harness reached on the corresponding model across environments. "Failed cells" counts environments where the harness scored exactly zero or produced no valid result, out of seven total environments.

Harness	Claude Code	Codex	OpenHands	A-Evolve	DeerFlow	ApodexHarness
Opus-4.8 mean	0.583	—	0.578	0.569	0.534	0.611
Opus-4.8 best	0.800	—	0.750	0.850	0.821	0.850
GPT-5.6-sol mean	—	0.559	0.560	0.648	0.580	0.592
GPT-5.6-sol best	—	0.900	0.950	0.900	0.800	0.950
Failed cells (Opus-4.8)	0	—	0	0	0	0
Failed cells (GPT-5.6-sol)	—	0	0	0	0	0

these two models' means only marginally, so it does not materially change the ordering between them.

Two patterns emerge from the table. First, best-cell scores cluster tightly across models, most exceeding 0.78, even though mean scores and failure counts vary substantially. This suggests that run-to-run consistency and harness compatibility, more than peak reasoning capability, drive much of the mean-score gap. Best-cell values are drawn from an asymmetric set of harness and environment pairings and are sensitive to noise and to how well a given harness matches a given model, so a definitive ceiling comparison requires paired analysis within fixed configurations. Second, the overall performance frontier remains narrow. Even the strongest configurations leave substantial room for improvement, and harness optimization mainly improves reliability rather than peak performance.

Harness performance: per-model and cross-model comparisons. Table 11 shows that no harness dominates once the model is held fixed: ApodexHarness leads on Opus, A-Evolve leads on GPT, and the ranking of the four model-agnostic harnesses differs substantially between the two model configurations. Harness effects and model effects are therefore both present, and neither is separable from the other using Claude Code or Codex, since these packaged harnesses are tightly coupled to a single model and cannot provide cross-model comparison. The four model-agnostic harnesses are architecturally compatible with all six models, but based on the two representative configurations tested here, no universal ranking of harness strength holds independent of model choice, and significant model–harness interaction is evident from the reversed rankings alone. Across all six foundation models, average performance of the four model-agnostic harnesses is led by ApodexHarness (0.548), followed closely by A-Evolve (0.544), then DeerFlow (0.521) and OpenHands (0.503). No single harness is universally optimal, and relative rankings shift substantially across model families.

Across the two representative closed-source models, ApodexHarness stays in the top half of the four model-agnostic harnesses on both; A-Evolve leads on GPT but drops to third on Opus, while OpenHands ranks second on Opus but falls to the bottom of the four model-agnostic harnesses on GPT. DeerFlow also shows substantial variability, ranking last on Opus but third on GPT. We adopt OpenHands as the primary instrument for the episode-mechanism ablations that follow for two reasons: it is fully open-source and auditable, and

Table 12: Episode-mechanism ablations under corresponding run configurations.

Run configuration			Outcome score		
Environment	Harness	Model	Baseline	Skill guidance	Verifier guidance
`LLM-pretrain-data-probe`	Claude Code	Opus-4.8	0.563	0.518 ▼ 0.045	–
	OpenHands	Opus-4.8	0.442	0.557 ▲ 0.115	0.507 ▲ 0.065
	OpenHands	GPT-5.6-sol	0.534	0.730 ▲ 0.196	0.659 ▲ 0.125
`pretrain-data-filter`	Claude Code	Opus-4.8	0.485	0.441 ▼ 0.044	–
	OpenHands	Opus-4.8	0.446	0.488 ▲ 0.042	0.460 ▲ 0.014
	OpenHands	GPT-5.6-sol	0.215	0.542 ▲ 0.327	0.523 ▲ 0.308

Note. Colored triangles denote descriptive changes from the corresponding latest appendix baseline. Verifier guidance is evaluated only with OpenHands because the intervention requires an instrumentable harness control loop. `LLM-pretrain-data-probe` cells average six episodes; `pretrain-data-filter` cells average its two canonical instances. Each instance has one run, so observed run-to-run variance limits causal interpretation of small differences.

its control loop can be directly modified to implement interventions such as verifier-guided repair.

These quantitative patterns are reinforced by qualitative observations that aggregates alone do not reveal. DeerFlow's shortfall is concentrated in premature termination, narrating a plan rather than executing it, a property of its control policy rather than of the underlying model. Full per-environment breakdowns are reported in Appendix C.

Our ApodexHarness illustrates that harness design effects can persist regardless of the foundation model with which the harness is paired. It outperforms OpenHands when paired with both GPT (0.592 vs. 0.560) and Opus (0.611 vs. 0.578), providing a comparison unconfounded by model choice because both harnesses were tested on the same two models. This pattern indicates that harness design makes a meaningful contribution to performance, though a formal decomposition of main effects and interactions is needed to definitively quantify the relative impacts of harness design versus model choice.

4.3.2. Episode-mechanism ablations

We next hold the harness, model, environment, and budget fixed and vary a single episode-level mechanism. We study two such mechanisms in turn: initial skill guidance, supplied at episode initialization and active throughout the trajectory, and verifier guidance, which reviews the work before final submission and gives the solver an opportunity to improve it. Note that here the "verifier guidance" should *not* be confused with "verifier-repair" loops presented in Sec. 4.4, because it is a lightweight in-environment guidance rather than an agentic system that provides diagnostics and repair guidance based on the entire solver trajectory. Table 12 reports the completed comparisons across three harness–model configurations. We use these runs as a focused analysis of mechanism behavior rather than as a broad estimate of average effects across tasks.

Skill guidance. The injected skill provides generic procedural guidance for a problem class; it is not tailored to any evaluated instance and contains no test items or solution information. Skill guidance is provided at episode initialization and shapes decisions throughout the trajectory rather than only at submission. On `LLM-pretrain-data-probe`, the OpenHands score increases from 0.442 to 0.557 with Opus-4.8 and from 0.534 to 0.730 with GPT-5.6-sol. The latter is the strongest reported condition and succeeds on all six episodes. By contrast, Claude Code–Opus-4.8 scores 0.518 with skill guidance, compared with its latest appendix baseline of 0.563. The same broad pattern holds on `pretrain-data-filter`: skill guidance improves both OpenHands configurations, most strongly with GPT-5.6-sol (+0.327), but not Claude Code–Opus-4.8

(−0.044).

Across both environments, the largest gains occur with OpenHands–GPT-5.6-sol, whereas Claude Code–Opus-4.8 has the highest unassisted baseline among the three configurations and does not improve with the injected skill. This may leave less room for generic guidance to add useful behavior, while differences in how a packaged harness incorporates the guidance may also affect how much of it reaches the trajectory. More broadly, skill guidance can partially compensate for weaknesses at the harness layer: although the unassisted OpenHands configurations start below Claude Code, their skill-guided scores match or exceed the Claude Code baseline on both tasks. Skill effectiveness should therefore be treated as a property of the complete solver configuration, and a well-matched skill can strengthen the overall system by narrowing a harness-level performance gap.

Verifier guidance. Verifier guidance instead inspects the work near submission and returns concrete issues for the solver to address before it is scored. The reviewer uses the same model and sees only the episode-visible task specification and transcript, with no access to hidden ground truth, reference answers, evaluation outcomes, or scores. It is also different from the process verifier to be discussed in Sec. 4.4, which is a much heavier system with stronger models to critique the process and trajectory of a system under testing. We evaluate this mechanism with OpenHands because its open-source control loop exposes the required intervention point. On `LLM-pretrain-data-probe`, verifier guidance increases the Opus-4.8 score from 0.442 to 0.507 and the GPT-5.6-sol score from 0.534 to 0.659. On `pretrain-data-filter`, the completed Opus-4.8 comparison changes from 0.446 to 0.460. The reviewer frequently identifies incomplete estimates and weak overlap measurements on `LLM-pretrain-data-probe`, indicating that outcome-blind feedback can locate deficiencies relevant to the final score even when it cannot inspect that score directly.

We select skill guidance and verifier guidance because they bracket the episode: skill shapes the trajectory from initialization, while verification intervenes near submission. In every completed OpenHands comparison, the guided score exceeds the corresponding appendix baseline, although the gain varies substantially by environment, model, and intervention. Skill guidance produces the larger gain in the configurations for which both interventions are complete. Together with the different result for Claude Code, this pattern shows that guidance does not contribute a fixed increment independent of the surrounding solver system. The reported differences remain descriptive because each instance has one run and the OpenHands prompt conditions are not fully controlled across the baseline and updated variants.

4.4. Verification-Driven Repair

Process verification yields a structured diagnosis rather than only a number, and that diagnosis can be returned to the solver to close a loop from measurement to improvement—a concrete and measurable discovery loop at the heart of discoverative AI. Because the subrubric-based HDS6 assessment assigns every band together with a citation to the trajectory steps that determine it, the assessment records both the location of each deficiency and its severity. From it we synthesize a *repair note*: a compact set of instructions that names the specific steps at fault, states the target band each should reach, and supplies a concrete, actionable correction for each. The note withholds the verified outcome, the hidden ground truth, and the outcome score, so that it guides the process without disclosing the answer. It is then inserted into the solver's prompt and the task is re-run under otherwise identical conditions, and re-scoring the repaired trajectory against the same rubric quantifies the effect of the guidance. Repair is in this sense a fixed procedure rather than a manual intervention, and it improves a trajectory without ever relaxing the isolation between solver and ground truth described in Section 3.1.

Table 13: Outcome effect of verification-driven repair on deficient trajectories (HDS6 < 3). For each environment we report the mean change in the environment's own outcome score between a trajectory and its repaired re-run, the number of trajectories, and how many improved, were unchanged, or declined.

Environment	Δ outcome	n	Improved	Unchanged	Declined
`solution_verifier`	+0.365	37	20	12	5
`aav_viability`	+0.360	34	21	7	6
`clinical_sap_tlf`	+0.271	90	41	48	1
`aav_capsid_enrichment`	+0.199	40	22	5	13
`LLM-pretrain-data-probe`	+0.082	51	27	3	21
`LLM_rollout_verifier`	+0.048	55	27	20	8
`posttrain_data_dedup`	+0.040	51	26	5	20
`internal_benchmark`	+0.034	28	7	19	2
`aav_sequence_design`	+0.010	33	9	19	5
`pretrain_data_filter`	−0.077	15	4	6	5
All environments	**+0.155**	**434**	**204**	**144**	**86**

We apply this procedure at scale. Every trajectory the process verifier places below band 3 on the HDS6 aggregate is paired with a repaired re-run of the same task under identical conditions, and the two are scored by the environment's own outcome verifier; the repair delta is the difference between them. Table 13 reports the mean delta per environment, over the environments whose outcome verifiers and required services were operational in this deployment. These results are experimental: the repair procedure is applied once to already-collected "deficient" trajectories (with HDS6 score < 3.0), and its effect on the environment's own outcome score is measured.

The effect is positive in nine of the ten environments and in the aggregate: across 434 deficient trajectories the repaired re-run scores 0.155 higher on average, with 204 trajectories improving against 86 declining. For trajectories the process verifier marks as deficient, the repaired re-runs improve the *outcome* on average and not merely the recorded process, and they do so from a diagnosis that never consults the outcome it improves. Because the study has no matched comparison group—the same deficient trajectories re-run without a repair note—we cannot fully separate this improvement from ordinary run-to-run variation.

In the rest of this section, we complement the results in Table 13 by presenting two case studies on how our HDS6 process verifier improves two concrete trajectories, from two episodes in different environments.

4.4.1. Case study: an under-specified clinical safety report

The `clinical_sap_tlf` environment poses a clinical-reporting task: reproduce a treatment-emergent adverse-event safety table from a statistical analysis plan (SAP). The plan is deliberately *under-specified*—it never names the patient population that forms the denominator, nor which adverse-event flag counts—so the solver must itself choose these analysis decisions and, ideally, declare them before computing, exactly the judgment a trial biostatistician is expected to exercise. Outcome scores incorporate both accuracy of the numerically calculated quantities, but also rigor, reproducibility and accountability of such calculations which are part of the report as submitted artifacts.

Figure 8 traces a matched pair of trajectories from the same solver (`claude-sonnet-5`) on one such episode: an initial run (left), the outcome-blind process critique of that run (centre), and the repaired re-run it induces (right). Both runs submit the same quantities as outcome, but their reports differ in terms of

rigor and accountability, as we will explain in the next paragraph. Reading only the first run's trajectory, the process critic flags two band-1 (of 2) subrubric failures. `boundary_and_scope` (the **Scope** capability) fires because the solver computed the table *before* declaring its under-specified population choice, leaving the assumptions log as after-the-fact justification; `claim_grounding` (the **Evidence** capability, here the subrubric of Table 2 instantiated for the clinical task) fires because the evidence ledger records the subject IDs behind each numerator but never behind the denominator, so the reported percentages cannot be independently recomputed.

The repair note carries these two findings—and only these—into the re-run. Guided by it, the repaired trajectory declares the population choice up front by interacting with the "virtual clinician" provided in the environment, and records patient IDs for both the numerators and the denominators in the submitted artifact, so that every cell becomes independently recountable. In terms of result, the first run's submitted artifact lists patient IDs only for the numerators and omits them for the denominators; its deliverable therefore cannot be fully recounted, and the outcome verifier scores it lower—0.83 against the repaired run's 0.95, a gap lying entirely on the evidence/traceability axis ($0.71 \rightarrow 1.00$). With the missing denominator ID lists supplied, both flagged subrubrics rise to band 2 and the outcome score rises accordingly.

4.4.2. Case study: a named estimator that was never run

The `LLM-pretrain-data-probe` environment poses a corpus-scouting task essential for data collection in LLM pre-training efforts: estimate how many unique high-value documents a collection of sources holds, together with their deduplicated token total, while drawing only small metered samples under a fixed budget. The brief states that naive document counting over-counts, because mirrors and near-duplicates re-serve the same content, and directs the solver to estimate the population by capture–recapture on the cross-source overlap. Because no source can be read in full, its size is never observed and must itself be inferred, so the submitted total rests on an inference the solver has to construct rather than on a quantity it can look up. The failure worth catching in this environment is therefore an estimate that carries the form of that inference without its substance, and it differs from the clinical case of Section 4.4 in a way that matters for repair: there the numbers were already correct and the deficiency lay in the audit trail, whereas here an ungrounded link propagates directly into the reported quantity.

Figure 9 traces a matched pair of trajectories from the same solver (`claude-sonnet-5`) on one episode of this environment: an initial run (left), the outcome-blind process critique of that run (centre), and the repaired re-run it induces (right). The initial run draws approximately thirty documents from each source and has an LLM-judge rate them, obtaining a high-value share of nine of thirty for `src_04`. It then supplies the remaining factor by assertion, taking each source to hold about one hundred documents, and scales the judged share by that constant, summing the per-source products together with a purchased 190-document package to submit $M_{\text{high}} = 108$ with the interval $[75, 145]$. The three crawlable sources in fact hold 451, 463 and 1498 documents, so the asserted constant is between four and fifteen times too small and the submitted total falls roughly an order of magnitude below the true value of 1505.

Reading only the initial trajectory, the process critic assigns `claim_grounding`—the **Evidence** capability, and the subrubric of Table 2—band 1 of 2, recording that the run named capture–recapture while computing a small-sample scale-up and that the content-hash fingerprints it had already downloaded went unused. The middle band, rather than the lowest, is the precise assessment, and the trajectory shows why: the sampling and the judging are genuine and are recorded step by step, and every arithmetic operation from the judged share to the submitted total is sound, so that exactly one link in the chain, the size of the source, rests on no observation. The integrity gate, which alone may consult the verified outcome, passes for both runs, which

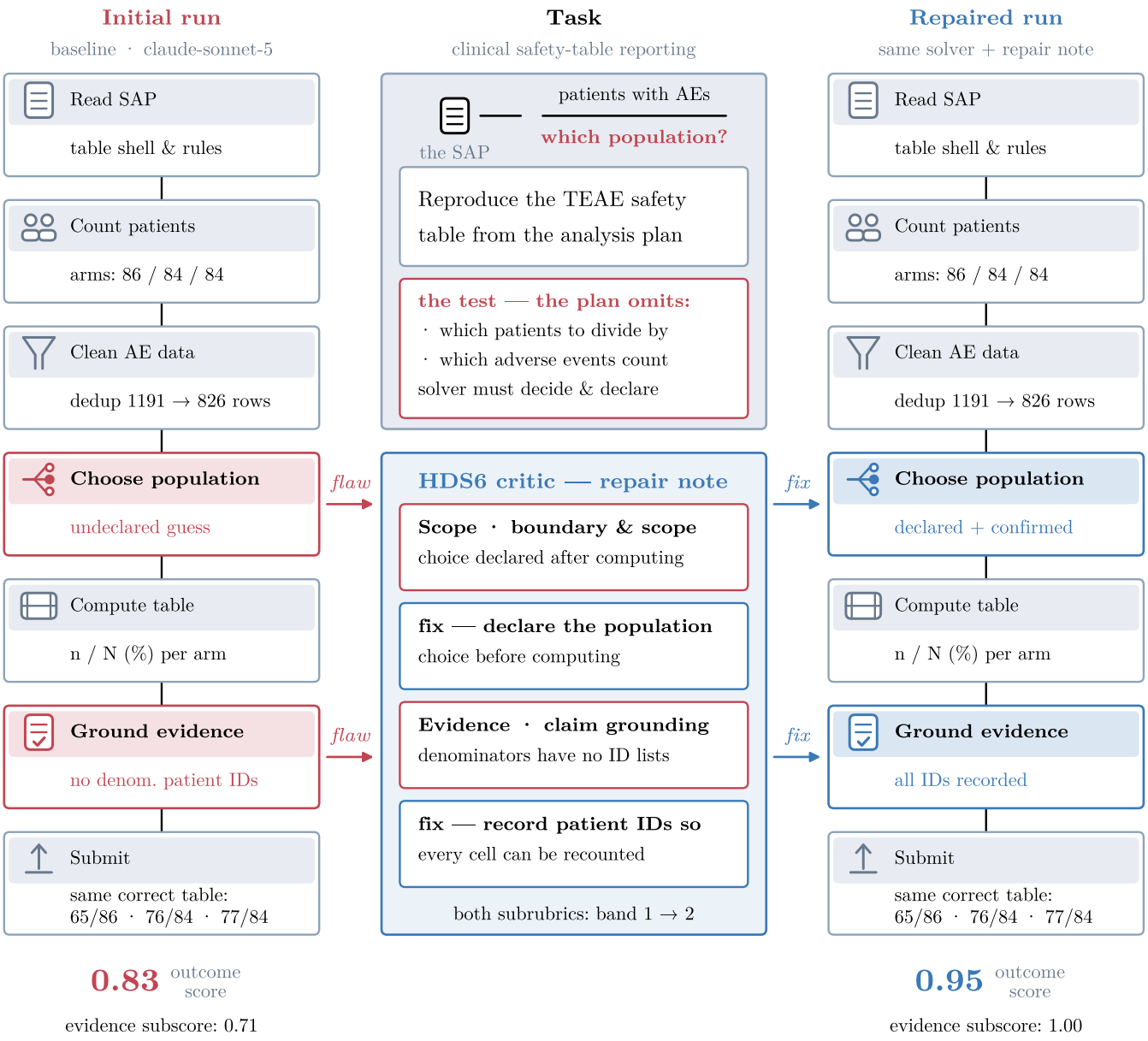

Figure 8: **A verification–repair episode on `clinical_sap_tlf`** (full walkthrough in Section 4.4). The task is to reproduce a treatment-emergent adverse-event (TEAE) safety table from the trial's statistical analysis plan (SAP); each cell divides the patients with adverse events (AEs) by a patient population that the plan deliberately never names — the fraction sketched in the task card. The left and right columns are the same solver's two attempts; white steps run identically in both, and the two red steps of the first attempt are where the critic's findings land. The centre column states the task and shows the outcome-blind HDS6 critique: each red-bordered box is a finding, each blue-bordered box the fix it prescribes, and the repair note carries both into the second attempt. The two runs submit the identical, correct table (bottom row), so the 0.83 → 0.95 score gap lies entirely on the evidence/traceability axis (0.71 → 1.00) which is part of the submitted artifact as reports and evidences of the said calculations.

places the deficiency in grounding rather than in fabrication. This configuration is the one an outcome score cannot resolve: a scalar records that 108 is wrong, whereas the band and its citation identify which input of the estimator was never measured, and that distinction is what makes the diagnosis actionable.

The repair note carries this finding into the re-run, directing the solver above all to use the per-document fingerprints it had already downloaded to measure cross-source overlap rather than to assume a size. Guided by it, the repaired trajectory re-samples each source and tabulates how often individual documents recur; for `src_04` it draws 450 times and observes 391 distinct documents, of which 336 appear exactly once and 51 appear exactly twice. Because first sightings dominate recurrences, most of the source remains unseen, and the Chao1 estimator accordingly returns $391 + 336^2/(2 \cdot 51) \approx 1498$ documents for that source in place of the one hundred previously asserted. Applying the same procedure to each source, discounting cross-source duplicates by a factor of 0.97 that the run set conservatively from the mirrors it had observed, and adding the purchased package yields $M_{\text{high}} = 2471$ with the interval $[2129, 2813]$.

The estimation score rewards how close the reported quantity is to the truth. The initial run's document and token counts are far from the true values, whereas the repaired run's are markedly closer, so its estimation score rises sharply—a clear gain on exactly the axis the critique named. The repaired estimate, while being closer to the truth in relative terms, still overshoots the truth, which places the remaining gap on a calibration axis this note did not target and marks interval-targeted repair as the next increment. The aggregate score improves as well, and its decomposition is instructive: procurement quality is essentially unchanged, but the extra measurement the correction requires consumes budget, so the efficiency term falls even as the estimate improves.

5. Related Work

Knowledge and reasoning benchmarks. A large body of work evaluates language models by comparing a single produced answer against a reference key. Broad knowledge and reasoning suites include MMLU [5], BIG-bench [26], the AI2 Reasoning Challenge [27], and the graduate-level GPQA [6], while mathematical ability is measured by GSM8K [28] and MATH [7]. More recent frontier suites such as FrontierMath [8] and Humanity's Last Exam [9] raise difficulty sharply while retaining the same answer-checking format. These benchmarks measure what a model knows and whether it can reason in a single pass, but they expose neither tools nor state, and they record nothing of the process by which an answer is reached.

Interactive and agentic benchmarks. A second line embeds a model in an executable environment and scores task success. The practice of evaluating agents by having them act in an environment has long-standing roots in reinforcement learning, from the Arcade Learning Environment [29] to OpenAI Gym [30]. This paradigm was later taken up by interactive environments for language agents, including the text-embodied world of ALFWorld [31], grounded web interaction in WebShop [32] and Mind2Web [33], and interactive coding with execution feedback in InterCode [34]. Software engineering is the most developed domain, spanning HumanEval [35], SWE-bench and its human-validated subset SWE-bench Verified [10, 36], and SWE-Gym [37]. Broader agentic suites target the web, the operating system, and general tool use, including WebArena [11], VisualWebArena [38], OSWorld [12], AppWorld [39], GAIA [40], AgentBench [41], τ-bench [13], ToolLLM [42], and Terminal-Bench [43], and specialized settings evaluate cybersecurity [44] and machine-learning or research engineering, as in MLAgentBench [45], MLE-bench [46], RE-Bench [47], and PaperBench [48]. Closest to the scientific domains we target are ScienceAgentBench [49], DiscoveryBench [50], the interactive science environment ScienceWorld [51], and LAB-Bench [52]. A complementary strand grounds tasks in measurable real-world value: GDPval [53] assembles professional deliverables across

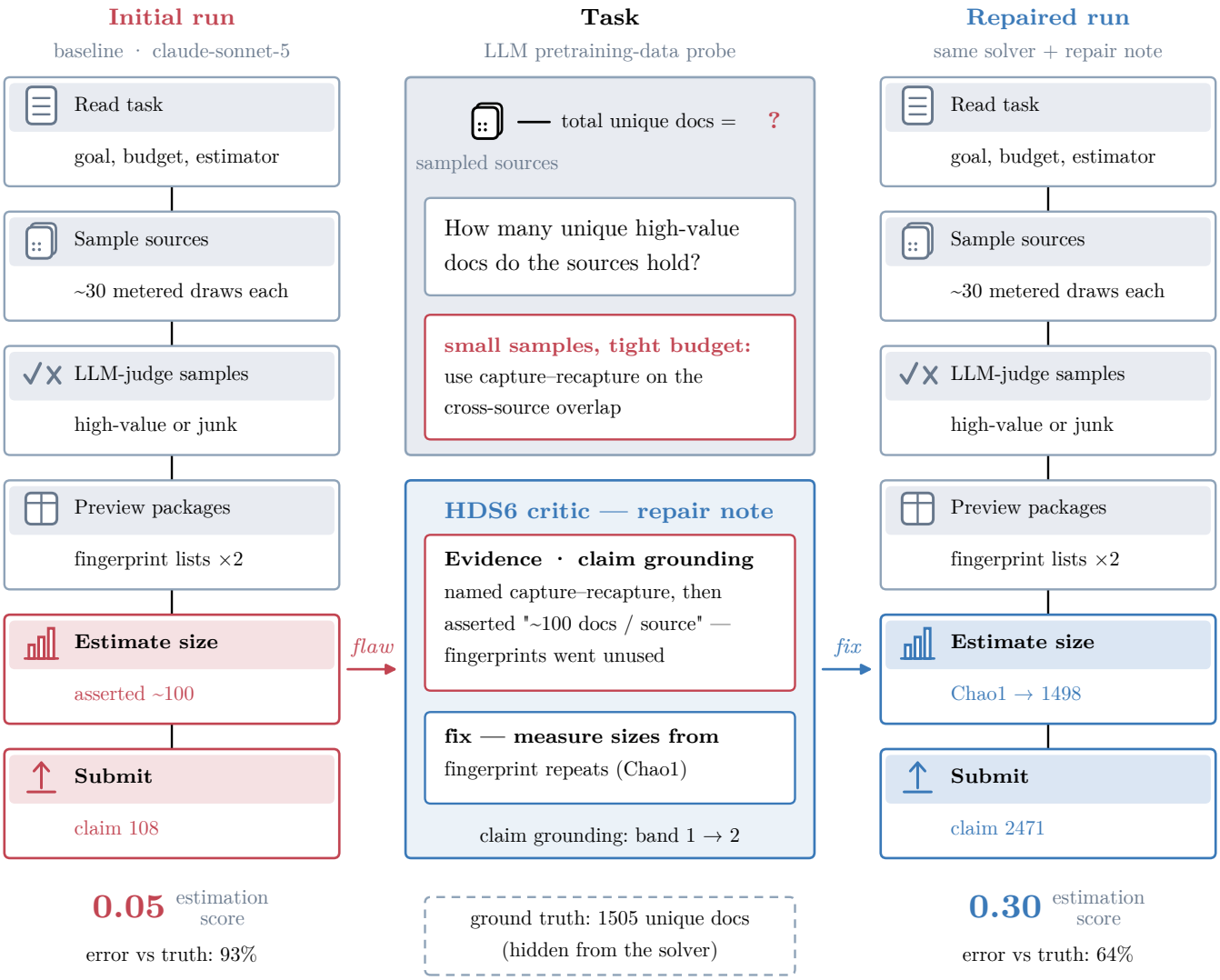

Figure 9: **A verification–repair episode on `LLM-pretrain-data-probe`** (full walkthrough in Section 4.4.2). The left and right columns are the same solver's two attempts at estimating how many unique high-value documents a corpus holds; white steps run identically in both. The first attempt asserts a source size of ∼100 documents and submits 108; the critic's finding (red-bordered box) records that the run named capture–recapture while the content-hash fingerprints it had downloaded went unused, and the prescribed fix (blue-bordered box) drives the repaired run to measure each source from its fingerprint repeats (Chao1 → 1498) and submit 2471. Against the hidden truth of 1505 documents (dashed box, never shown to the solver), the two claims are 93% and 64% off, and the estimation score moves 0.05 → 0.30.

dozens of economically significant occupations, SWE-Lancer [54] draws software tasks from real freelance contracts with associated payouts, and Agents' Last Exam [55] evaluates agents on existing jobs that human professionals perform today, while the time-horizon analysis of Kwa et al. [56] measures the length of tasks that agents complete reliably. These benchmarks establish interactivity and real-world value across an expanding range of domains, and TRACES builds on them by verifying load-bearing intermediate results and reporting a process-quality signal alongside task success, within a single environment–task–episode abstraction that spans biomedical, clinical, and frontier-model-engineering problems.

Agent scaffolds. Because we evaluate a model together with the harness that drives it, the scaffolds that turn a model into an agent are directly relevant. General prompting and control strategies include ReAct [57], Tree of Thoughts [58], Reflexion [59], and Self-Refine [60]; tool use is enabled by Toolformer [61]; and multi-agent and open-ended frameworks include AutoGen [62] and Voyager [63]. In software engineering, agent–computer interfaces such as SWE-agent [64] and platforms such as OpenHands [24] package a model into a complete solver. Iterative self-correction has been studied both as a tool-grounded procedure, as in CRITIC [65], which revises outputs through interaction with external tools, and as an intrinsic capability, where Huang et al. [66] find that self-correction without external feedback often does not improve, and can degrade, reasoning; these findings motivate the verifier-grounded feedback on which our repair loop depends. Our framework treats the surrounding scaffold as part of the system under evaluation rather than as fixed infrastructure, and it therefore measures a model and its harness jointly.

Automated and LLM-based evaluation. Using strong models to judge the outputs of others has become standard practice, from pairwise, arena-style judging [67] to reference-free scoring with G-Eval [68] and open evaluator models such as JudgeLM [69] and Prometheus 2 [70]. Rubric- and skill-based protocols, including FLASK [71] and, in medicine, HealthBench [72], decompose quality into explicit criteria. Closest to process evaluation, trained critics such as CriticGPT [73] produce natural-language critiques that surface errors in model-written code and can match or exceed human reviewers at locating them. These judges are typically single-pass, outcome-facing, and aware of the model identity or the reference answer. Our process verifier differs on each count: it is agentic rather than single-pass, it grounds every judgment in cited trajectory evidence, and it is blind both to the verified outcome and to the identity of the solver.

Process supervision and verifiable rewards. A related thread supervises the reasoning process rather than only the final answer. Uesato et al. [74] contrasted process- and outcome-based feedback on mathematical problems, and Lightman et al. [75] then showed that step-level supervision produces markedly stronger verifiers; Math-Shepherd [76] and generative verifiers [77] subsequently reduced the annotation cost of such process reward models, and the reliability of step-level judgments has itself become an object of study, as in ProcessBench [78], which measures how accurately a critic localizes the first erroneous step in a reasoning trace. In parallel, reinforcement learning from verifiable rewards has driven recent reasoning models, including DeepSeek-R1 [79] and open recipes such as Tülu 3 [80], building on the verifier-based approach to math word problems of Cobbe et al. [28]. Our HDS6 metric extends process evaluation from short arithmetic chains to long, tool-using trajectories, and grades them against task-specific, capability-level rubrics rather than a single scalar reward.

Specification gaming and contamination. Our reliance on hidden verifiers and anti-gaming gates connects to work on specification gaming and reward hacking [81, 82] and to the growing concern over benchmark data contamination [83, 84]. Both literatures motivate verifiers that a solver cannot inspect and outcomes that cannot be reached by shortcut, which are design commitments of the executable environment.

Autonomous scientific discovery. Finally, a growing body of systems pursues end-to-end scientific research. Coscientist [1] and ChemCrow [85] couple large language models to laboratory tools and automation to plan and execute chemistry experiments; the AI co-scientist [2] generates and refines novel research hypotheses through multi-agent debate; The AI Scientist [3] carries out machine-learning research end to end; and broader analyses chart the emerging role of biomedical AI agents [86]. More broadly, AI is increasingly central to scientific discovery [87], with concrete results ranging from new algorithms found by evolutionary coding agents [88] to the large-scale discovery of stable materials [89] and the autonomous laboratories that synthesize them [90]. These systems share our long-term goal of open-ended discovery, and Apodex Discovery supplies the reality-based environments, hidden verifiers, and process-level evaluation through which such systems can be measured, compared, and improved on consequential problems.

6. Conclusion

Apodex Discovery advances three complementary capabilities for building and evaluating discoverative AI on consequential real-world problems. It provides a systematic process for identifying problems that are both high-value and objectively verifiable; it converts those problems into executable environments in which complete solver systems can act through data, tools, and feedback; and it introduces HDS6 to evaluate the quality of the investigation itself when definitive outcome ground truth is delayed, incomplete, or not yet available. Together, these components shift evaluation from asking whether a model produced the right answer to asking whether a solver conducted a reliable, evidence-grounded, and self-correcting investigation. Alongside this paper we release **TRACES**, the reality benchmark that instantiates this framework: seventeen executable environments spanning 218 episodes across biomedicine, clinical translation, and frontier-model engineering, all with hidden verifiers and two supporting the full verification–repair loop.

Our initial results show why this broader evaluation unit matters. In adeno-associated virus capsid design, domain-specific environments improved performance across the discovery pipeline, with Apodex exceeding prior human baselines on the most demanding tasks. In drug repurposing, the biomedical environment improved the mean normalized prediction score of GPT-5.5 and GPT-5.6-sol by 2.5 and 7.6 points, respectively, over the same closed-book backbone. Controlled ablations further demonstrate that Apodex Discovery can attribute performance differences to distinct parts of the solver system—including the foundation model, harness, guidance, budget handling, and verifier-driven repair—rather than treating every failed episode as a generic model-capability failure.

Limitations and Extensions. The TRACES benchmark currently instantiates seventeen executable environments with hidden verifiers, spanning 218 episodes and realizing a subset of the twenty selected high-value problems across biomedicine, clinical translation, and frontier-model engineering; two of these environments support the full verification–repair loop. This already constitutes a substantial benchmark, although coverage remains limited to this initial subset of the 423-problem registry, with many additional environments under development and expected to be released in the near future. Our process-verification results likewise provide a strong initial demonstration while leaving clear room for further development, to align process verification scores better with outcome scores and demonstrate its effectiveness over an even broader range of applications.

7. Author Contributions and Acknowledgments

We gratefully acknowledge Mr. Tianqiao Chen, the project lead of this work, who proposed the concept of the heavy-duty solver, defined the six capabilities of the HDS6 process metric, and shaped the overall

architecture of Apodex Discovery. We sincerely thank him for the vision, guidance, and support that made this work possible.

We would like to acknowledge Dr. Yuan Cai, Dr. Chengyu Fu, Dr. Shiyi Zhang, and Harry Zhang for the insightful discussion and contribution.

Core technical contributors. Sheng Wang, Brian Wang, Bin Feng, Xiaoman Pan, Chenyang An, Felix Liu.

Contributors. Tangqi Fang, Gongbo Sun, Yueqi Guo, Kailong Wen, Feng Xing, Yiling Guo, Lingfeng Shen, Ning Wang, Handuo Zhang, Feng Chen, Fuchao Yang, Jiacheng Lin, Siting Li, Zixuan Liu, Xiang Wang, Chi Han, Zhenhailong Wang, Kunlun Zhu, Lawrence Zhao, Lidong Bing, David Tan, Bo An, Heng Ji.

References

[1] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[2] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[3] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[4] Jeff Dean, Sanjay Ghemawat, Oriol Vinyals, and Quoc Le. Discovery loop. https://www.discoveryloop.com/, 2026. Public benefit corporation for automating the experimental loops of science and engineering.

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *Conference on Language Modeling (COLM)*, 2024.

[7] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

[8] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. arXiv preprint arXiv:2411.04872, 2024.

[9] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, et al. Humanity's last exam. arXiv preprint arXiv:2501.14249, 2025.

[10] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024.

[11] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

[12] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[13] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

[14] Drew H. Bryant, Ali Bashir, Sam Sinai, Nina K. Jain, Pierce J. Ogden, Patrick F. Riley, George M. Church, Lucy J. Colwell, and Eric D. Kelsic. Deep diversification of an AAV capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

[15] Pierce J. Ogden, Eric D. Kelsic, Sam Sinai, and George M. Church. Comprehensive aav capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science (New York, N.Y.)*, 366:1139 – 1143, 2019. URL https://api.semanticscholar.org/CorpusID:208358760.

[16] Fatma-Elzahraa Eid, Albert T Chen, Ken Y Chan, Qin Huang, Qingxia Zheng, Isabelle G Tobey, Simon Pacouret, Pamela P Brauer, Casey Keyes, Megan Powell, et al. Systematic multi-trait aav capsid engineering for efficient gene delivery. *Nature Communications*, 15(1):6602, 2024.

[17] Qin Huang, Albert T. Chen, Ken Y. Chan, Hikari Sorensen, Andrew J. Barry, Bahar Azari, Qingxia Zheng, Thomas Beddow, Binhui Zhao, Isabelle G. Tobey, Cynthia Moncada-Reid, Fatma-Elzahraa Eid, Christopher J. Walkey, M. Cecilia Ljungberg, William R. Lagor, Jason D. Heaney, Yujia Alina Chan, and Benjamin E. Deverman. Targeting aav vectors to the central nervous system by engineering capsid–receptor interactions that enable crossing of the blood–brain barrier. *PLOS Biology*, 21, 2023. URL https://api.semanticscholar.org/CorpusID:259983952.

[18] Jason Wu, Yu Qiu, Eugenia Lyashenko, Tess Torregrosa, Edith L. Pfister, Michael J. Ryan, Christian Mueller, and Sourav R. Choudhury. Prediction of adeno-associated virus fitness with a protein language-based machine learning model. *Human Gene Therapy*, 36:823 – 829, 2024. URL https://api.semanticscholar.org/CorpusID:271915352.

[19] Binjie Guo, Hanyu Zheng, Xurong Lin, Haohan Jiang, Aisheng Mo, Hongyan Wei, Ru Zhang, Yubin Yuan, Yile Wu, Hengguang Li, et al. Mapping aav capsid sequences to functions through function-guided in silico evolution. *Cell Press Blue*, 1(3), 2026.

[20] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, 40(7):1114–1122, 2022.

[21] Abramson et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[22] Mohammadreza Ghaffarzadeh-Esfahani and Yousof Gheisari. Aavgen: Precision engineering of adeno-associated viral capsids for renal selective targeting. *ArXiv*, abs/2602.18915, 2026. URL https://api.semanticscholar.org/CorpusID:285974345.

[23] Lijun Liu, Jiali Yang, Jianfei Song, Xinglin Yang, Lele Niu, Zeqi Cai, Hui Shi, Tingjun Hou, Chang-yu Hsieh, Weiran Shen, et al. Aavdiff: Experimental validation of enhanced viability and diversity in recombinant adeno-associated virus (aav) capsids through diffusion generation. *arXiv preprint arXiv:2404.10573*, 2024.

[24] Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations (ICLR)*, volume 2025, pages 65882–65919, 2025.

[25] Minhua Lin, Hanqing Lu, Zhan Shi, Bing He, Rui Mao, Zhiwei Zhang, Zongyu Wu, Xianfeng Tang, Hui Liu, Zhenwei Dai, et al. Position: Agentic evolution is the path to evolving llms. *arXiv preprint arXiv:2602.00359*, 2026.

[26] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

[27] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[28] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[29] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[30] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

[31] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[32] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[33] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[34] John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. InterCode: Standardizing and benchmarking interactive coding with execution feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[35] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[36] OpenAI. Introducing SWE-bench verified. https://openai.com/index/introducing-swe-bench-verified/, 2024. Technical report.

[37] Jiayi Pan et al. Training software engineering agents and verifiers with SWE-Gym. *arXiv preprint arXiv:2412.21139*, 2024.

[38] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[39] Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[40] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

[41] Xiao Liu, Hao Yu, Hanchen Zhang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

[42] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations (ICLR)*, 2024.

[43] The Terminal-Bench Team. Terminal-Bench: A benchmark for AI agents in terminal environments. https://www.tbench.ai/, 2025.

[44] Andy K. Zhang et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*, 2024.

[45] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *International Conference on Machine Learning (ICML)*, 2024.

[46] Jun Shern Chan et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

[47] Hjalmar Wijk et al. RE-Bench: Evaluating frontier AI r&d capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.

[48] Giulio Starace et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

[49] Ziru Chen et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

[50] Bodhisattwa Prasad Majumder et al. DiscoveryBench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

[51] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[52] Jon M. Laurent et al. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

[53] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

[54] Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-Lancer: Can frontier LLMs earn $1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*, 2025.

[55] Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026.

[56] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, et al. Measuring AI ability to complete long software tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[57] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[58] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[59] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[60] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[61] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[62] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[63] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[64] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[65] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*, 2024.

[66] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

[67] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[68] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[69] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

[70] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[71] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations (ICLR)*, 2024.

[72] Rahul K. Arora et al. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[73] Nat McAleese, Rai Michael Pokorny, Juan Felipe Cerón Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM critics help catch LLM bugs. *arXiv preprint arXiv:2407.00215*, 2024.

[74] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[75] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

[76] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[77] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[78] Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[79] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[80] Nathan Lambert et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[81] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[82] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[83] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

[84] Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[85] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

[86] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151, 2024.

[87] Hanchen Wang, Tianfan Fu, Yuanqi Du, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[88] Alexander Novikov et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[89] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[90] Nathan J. Szymanski et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

[91] Sankar Basu and Björn Wallner. Dockq: A quality measure for protein-protein docking models. *PLoS ONE*, 11, 2016. URL https://api.semanticscholar.org/CorpusID:11540847.

[92] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

[93] John M. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

[94] Ezra J. Loeb, Sophia A. Fergione, Vivian Yudistyra, Marco M. Fanous, Abigail R. Benkert, Delaney G. Fisher, Joshua A Hull, Mai K. ElMallah, and Aravind Asokan. Complete neutralizing antibody evasion by serodivergent non-mammalian aavs enables gene therapy redosing. *Cell Reports Medicine*, 6, 2025. URL https://api.semanticscholar.org/CorpusID:283352205.

[95] Qiaozhen Meng, Fei Guo, Ercheng Wang, and Jijun Tang. Comdock: A novel approach for protein-protein docking with an efficient fusing strategy. *Computers in biology and medicine*, 167:107660, 2023.

A. Domains and questions considered by Apodex Discovery

We list below the domains and questions considered by Apodex Discovery.

Table 14: The ten selected domains.

#	Domain	Why the value is high	Why it suits a long-chain reasoning agent	Representative players
1	Bioinformatics, Computational Biology & Genomic Medicine *Healthcare & Life Sciences*	Diagnosis-to-therapy genomics approaches a $100B vertical; breakthrough algorithmic biology carries near-Nobel recognition.	The agent assembles alignment and variant-calling pipelines, runs them, and verifies output against held-out truth sets; fuses multi-omics layers and reasons over pathways to testable hypotheses with quantified uncertainty.	No single incumbent — an AI-startup-saturated category; pharma R&D informatics contracts limit the flywheel.
2	Foundation Models for Scientific Reasoning *Software & Internet*	A true scientific-reasoning foundation model would be Turing/Nobel-class *and* a >$100B platform under all of R&D — it maxes both axes.	The agent drafts a proof, formalises it in Lean and accepts only what the kernel verifies; derives governing equations, discretises them to simulator code, and cross-checks against numerical ground truth.	Frontier labs — the single hottest AI-native battleground; model-swap risk weakens lock-in.
3	Scientific HPC Code Generation & Optimization *Software & Internet*	Developer-productivity and compute, a $10–100B opportunity adjacent to the AI-coding boom.	The agent ports CUDA to HIP/SYCL and verifies bitwise numerical agreement; applies tiling and vectorisation, measures speedup, and rejects regressions at a CI benchmark gate.	Copilot, Cursor and the hyperscalers contest general code generation; the scientific-HPC niche is less contested.
4	Healthcare IT & Digital Health *Healthcare & Life Sciences*	EHR, telehealth and monitoring form a vertical approaching $100B with platform-scale winners.	Long-horizon reasoning over longitudinal records, with outcomes that arrive as data rather than as opinion.	Epic and the incumbent EHR platforms.
5	Weather-Driven Commodity & Energy Market Forecasting *Agriculture & Food (+ Energy & Environment)*	Grain and soft commodities each exceed $100B, dominated by a few majors controlling sourcing, storage and logistics; energy trading supports a >$100B opportunity.	Physical-world forecasting where the verifier is the official report that lands weeks later — a genuine held-out future, not a stored key.	The ABCD grain majors; concentrated merchant and roaster franchises.
6	Structural & Civil Engineering Services *Real Estate & Construction*	A $10–50B global vertical, squarely in the $10–100B band.	Code compliance is machine-checkable: layouts can be generated, then verified against the standard and against seismic constraints.	Legacy CAD incumbents; concentrated engineering firms make the buyer reachable.
7	Chip Design & Verification (EDA) *Hardware & Semiconductors*	A $10–100B concentrated, high-margin tooling industry essential to all chip design; the fabless model it serves is a >$100B market.	The whole design pipeline runs inside the tools and is verified *in silico* before fabrication — a fast, exact loop over a hard search.	Synopsys and Cadence in signoff tooling; TI and ADI in analog; Nvidia, Qualcomm, Broadcom, AMD as fabless designers.

continued on next page

Table 14 — continued from previous page

#	Domain	Why the value is high	Why it suits a long-chain reasoning agent	Representative players
8	Market Data & Financial Indices *Financial Services*	Oligopolistic, extremely high-margin franchises whose aggregate addressable revenue exceeds $100B.	Claims are settled by the market itself on a fixed horizon; out-of-sample tracking is native to the domain.	Bloomberg, S&P, MSCI.
9	Materials Discovery, Design & Qualification *Materials & Chemicals*	The materials-discovery prize, with recognition on the scale of the Materials Genome effort, over a $10–100B-plus advanced-materials market.	Inverse design closes a loop: propose a structure, predict the property, and qualify against measurement.	Crowded with AI-native entrants; DeepMind's GNoME contests it at hyperscaler scale.
10	Industrial Automation & Control *Industrials & Manufacturing*	PLC, DCS and MES software and hardware form a vertical approaching $100B — critical infrastructure with no public fame.	Control problems are formalisable and simulatable, with plant telemetry as the verifier.	Siemens, Rockwell, Emerson, Schneider.

Table 15: The twenty-question registry.

#	Question	What the solver must do	How the answer is checked	Buyer
1	LLM training recipe design *Frontier model R&D*	Design and optimise pretrain / midtrain / posttrain recipes — architecture, optimiser, schedule, RL, distillation — under a compute constraint.	Must beat expert-tuned baselines at matched compute.	Frontier and enterprise model builders
2	Training data curation & decontamination *Frontier model R&D*	Discover, filter, construct and decontaminate training and benchmark data at scale, under budget and throughput limits.	Downstream quality at fixed budget; contamination measurably removed.	Frontier and enterprise model builders
3	LLM training & serving infrastructure *Frontier model R&D*	Build and verify reliable, high-throughput training and serving infrastructure across models and hardware.	Bit-exact reproduction and measured throughput across targets.	Model builders; compute providers
4	Clinical trial statistical programming *Healthcare & Life Sciences*	Generate SAP-faithful trial analysis code and top-line tables, listings and figures.	Cell-level audit lineage back to the protocol; fidelity to the statistical analysis plan.	Pharma biostatistics; CROs
5	Clinical trial safety & efficacy signal detection *Healthcare & Life Sciences*	Auto-generate and stress-test safety and efficacy hypotheses from accruing trial data.	Signals validated against post-cutoff outcomes.	Pharma; medical monitors; CROs
6	Drug repurposing and reformulation *Healthcare & Life Sciences*	Rank repurposing candidates for a target disease using only evidence visible before a cutoff.	Enrichment of the top-K against post-cutoff clinical and real-world success.	Pharma; biotech; disease foundations
7	Causal drug-target validation *Healthcare & Life Sciences*	Decide whether a candidate gene target is a causal driver or an associated passenger, and name the falsifying experiment.	Held-out genetics and perturbation outcomes; known target successes and failures.	Pharma discovery; biotech

continued on next page

Table 15 — continued from previous page

#	Question	What the solver must do	How the answer is checked	Buyer
8	Non-consensus investment thesis generation *Financial Services*	Surface overlooked evidence, non-consensus theses, and bull / base / bear scenarios for deep-tech decisions.	Historical time-split replay, shadow portfolio, out-of-sample thesis tracking.	Asset managers; venture and growth investors
9	Crop yield shortfall forecasting *Agriculture & Food*	Predict localised crop-yield shortfalls sixty days ahead of official reports, from satellite, soil and climate data.	The official report, when it lands.	Agricultural traders; food processors; insurers
10	Code-compliant structural design *Real Estate & Construction*	Generate structural and foundation layouts from a floor plan, minimising embodied carbon under seismic constraints.	Automated code-compliance check; carbon and seismic constraints evaluated.	Engineering and design firms; developers
11	Cell therapy manufacturing optimization *Healthcare & Life Sciences*	Optimise per-batch manufacturing parameters to raise potency and post-infusion persistence.	Measured batch potency and persistence after infusion.	Cell-therapy manufacturers; CDMOs
12	AAV capsid assessment and design *Healthcare & Life Sciences*	Design organ-specific AAV capsid variants end-to-end from public sequence–function data.	Unseen variants under a leak-proof hold-out.	Gene-therapy developers
13	Biological age & mortality risk prediction *Healthcare & Life Sciences*	Estimate biological age and mortality risk from multimodal health data — records, imaging, labs.	Unseen held-out cohorts.	Health systems; insurers; longevity clinics
14	iPSC-cardiomyocyte differentiation modeling *Healthcare & Life Sciences*	Model iPSC-to-cardiomyocyte differentiation trajectories and the effect of disease mutations.	Unseen donors and unseen mutations.	Cardiac drug discovery; academic labs
15	Rare disease diagnosis from whole-genome data *Healthcare & Life Sciences*	Produce structured diagnostic reports from raw whole-genome or whole-exome data and match patients to trial eligibility.	Confirmed diagnoses; eligibility verified against trial criteria.	Rare-disease centres; patient organisations
16	Metabolic treatment response prediction *Healthcare & Life Sciences*	Predict response and non-response, and rank the best twelve-month interventions per patient.	Twelve-month weight, HbA1c, discontinuation and rebound outcomes.	Payers; health systems; pharma
17	Preclinical-to-human translation prediction *Healthcare & Life Sciences*	Predict whether a preclinical result will translate, and attribute the cross-species causal gap when it does not.	Paired preclinical and human outcomes on held-out drugs.	Pharma R&D; biotech
18	Near-term health deterioration prediction *Healthcare & Life Sciences*	Predict near-term health decline from longitudinal personal data and recommend safe next actions.	Observed near-term decline and escalation outcomes.	Health systems; payers; care providers
19	Work-state reconstruction & open-loop detection *Software & Internet*	Reconstruct active work state across a company's tools and detect unresolved commitments, decisions and risks.	Whether flagged items were genuinely open, and whether the loop closed.	Enterprises
20	Material-event detection from real-time news *Software & Internet*	Monitor real-time signals for events that materially change a user's world model, with alternative readings.	Whether flagged events proved material, judged after the fact.	Enterprises; investors; analysts

B. Adeno-associated virus capsid design full results

Task 1: Viability. Each variant is classified as viable (it assembles and packages its genome) or not, from its amino-acid sequence alone. The two instances test complementary extrapolation, a multi-mutation instance that extrapolates along mutational load [14], and a single-mutation instance tiled across the whole capsid sequence that extrapolates across sequence position and mutation type [15]. Because viability is strongly confounded with mutational load, a predictor that merely counts mutations already scores well; the primary metric is therefore the AUROC computed separately within each exact mutation count and then averaged, isolating whether a model ranks survivors above non-survivors among equally mutated variants (chance 0.5). The multi-mutation instance trains on variants with at most 12 mutations and tests on those with 13 or more. It has 139,268 training and 133,433 test variants, plus 20,816 behind the validation oracle. The single-mutation instance trains on substitutions and tests on insertions and deletions. It has 15,409 training and 13,891 test variants, plus 1,544 behind the oracle.

Task 2: Tropism. A seven-amino-acid peptide inserted into a fixed surface loop of the AAV9 capsid is scored for its ability to reach each target tissue, against the measured enrichment of each peptide in that tissue from the Fit4Function screens [16]. Three instances of increasing difficulty are reported, multi-organ tropism in mouse, cross-species transfer from mouse and in-vitro data to macaque, and human in-vitro assays. The primary metric is the rank correlation between predicted and measured enrichment (Spearman for the mouse and cross-species instances, mean per-assay Pearson for the human in-vitro instance), The mouse instance draws 73,089 training and 15,679 validation peptides at random from one library and is scored on 6,032 held-out ones. Every test peptide there is at Hamming distance at least 2 from every training peptide. The human in-vitro instance splits its own library the same way: 66,476 training, 14,244 validation and 14,244 test peptides. The cross-species instance trains on 50,235 mouse and in-vitro peptides and is scored on 13,319 peptides from a separate macaque library, with 12,589 more for validation.

Task 3: Structure. Given only the amino-acid sequence, a solver must produce the three-dimensional capsid structure by orchestrating a provided set of folding engines and analysis tools inside a network-isolated environment. Three structural levels are scored, the variable surface loops of the single subunit, by per-residue positional accuracy on the loop atoms; the full 60-mer icosahedral shell, by a geometric-accuracy term and a physical-plausibility term; and the capsid in complex with an antibody or receptor, by an interface-docking quality [91] and the fraction of recovered native interface contacts. Leakage is controlled by a temporal split (only structures deposited after 30 September 2021, past each folding engine's training cut-off), sequence-similarity filtering, and no network access. The loop and shell levels score the same 12 post-cutoff capsids. Each is under 95% identical to its closest match among the 62 pre-cutoff structures released as a reference pool. The complex level scores 15 targets: 8 with an antibody and 7 with the AAVR receptor.

Task 4: Design. A solver generates up to 1000 novel peptides for a target. It starts from a labeled seed set: 2,000 training and 500 validation peptides on each tissue instance, 188,000 and 10,000 on each receptor instance. It may train surrogate models on that set. It may also query a scoring oracle, but on at most 100 sequences, a tenth of what it submits, so the oracle cannot be brute-forced. A design counts only if it is simultaneously novel (at least two substitutions from every training peptide), manufacturable (passing a viability gate), and on target. On-target is judged by rank, for the tissue instances (brain, spinal cord, heart), a design qualifies if its predicted organ selectivity ranks in the top 1%, 2%, or 5% of all sequences, and the instance score is the pass rate averaged over these thresholds; for the receptor instances (LY6A, LY6C1), if a trained classifier judges it specific. The oracles are trained on real fitness data, the Fit4Function screens [16] and LY6A/LY6C1 binding measurements [17].

Table 16: Viability prediction. Tested on two held-out instances, with validation labels hidden and reachable only through a metered scoring action. The score is out-of-distribution ranking accuracy, the area under the ROC curve for separating viable from non-viable variants, averaged over different number of mutations. Scores are means over valid repeats.

	Multi-mutation		**Single-mutation**		
model	score	HDS6	score	HDS6	**mean**
kimi-k3	**0.962**	3.48	0.910	3.51	0.936
claude-opus-5	0.959	3.43	**0.913**	3.49	0.936
glm-5.2	0.960	3.37	0.905	3.18	0.932
claude-opus-4-8	0.951	**3.62**	0.912	**3.68**	0.931
claude-sonnet-4-6	0.952	3.36	0.893	3.55	0.922
Claude Code (opus 4.8)	0.950	—	0.891	—	0.921
deepseek-v4-pro	0.941	3.33	0.894	3.41	0.918
gpt-5.5	0.927	2.85	0.891	3.14	0.909
apodex-1.1	0.931	3.27	0.877	3.44	0.904
qwen3.5-397b	0.890	3.24	0.902	3.31	0.896
apodex-1.0	0.905	3.17	0.863	3.30	0.884
CAP-PLM [18] (published SOTA)	0.897	—	0.859	—	0.878
ALICE-GB [19]	0.885	—	0.696	—	0.791
augmented ridge [20]	0.874	—	0.848	—	0.861
logistic additive [14]	0.888	—	0.852	—	0.870

Table 17: Tropism prediction. Seven-amino-acid inserts in a fixed surface loop of the AAV9 capsid, scored by the agreement between predicted and measured per-target enrichment across mouse multi-organ tropism (A), cross-species transfer from mouse and in-vitro data to macaque (B), and human in-vitro assays (C). Scores are per-target means over valid repeats.

	A · Mouse tropism		**B · Cross-species**		**C · In-vitro cell**		
solver	score	HDS6	score	HDS6	score	HDS6	**mean**
claude-opus-5	**0.562**	3.65	**0.638**	3.10	**0.760**	**3.90**	0.653
kimi-k3	0.559	3.70	0.637	3.63	0.751	3.51	0.649
glm-5.2	0.553	**3.73**	0.626	**3.84**	0.752	3.72	0.644
claude-opus-4-8	0.548	3.70	0.619	3.40	0.740	2.80	0.635
apodex-1.1	0.539	3.16	0.621	3.57	0.746	3.37	0.635
Claude Code (opus 4.8)	0.545	—	0.617	—	0.738	—	0.633
claude-sonnet-4-6	0.549	3.42	0.610	1.91	0.725	2.97	0.628
deepseek-v4-pro	0.508	3.23	0.602	2.81	0.728	2.88	0.613
qwen3.5-397b	0.500	2.98	0.600	2.84	0.716	2.96	0.605
gpt-5.5	0.534	3.56	0.575	3.75	0.691	3.39	0.600
apodex-1.0	0.525	2.43	0.540	2.64	0.489	3.62	0.518
F4F[16] (published SOTA)	0.532	—	0.614	—	0.719	—	0.622

Table 18: Structure prediction. The variable surface loops of a single subunit, the full 60-mer icosahedral shell, and the capsid in complex with an antibody or receptor. Scores measure agreement with the reference structure, per-residue positional accuracy for the surface loops and the shell, and interface quality for the complex. Scores are means over valid repeats.

	Loop		**Assembly**		**Complex**		
solver	score	HDS6	score	HDS6	score	HDS6	**mean**
kimi-k3	0.924	**3.56**	0.754	3.60	**0.359**	**3.89**	0.679
claude-opus-5	0.913	3.43	0.768	3.79	0.322	3.60	0.668
glm-5.2	0.922	3.31	0.736	3.53	0.336	3.33	0.665
claude-opus-4-8	0.927	3.28	**0.769**	**3.82**	0.275	2.99	0.657
apodex-1.1	**0.928**	3.47	0.742	3.60	0.277	3.49	0.649
deepseek-v4-pro	0.927	2.50	0.749	2.94	0.154	1.13	0.610
Claude Code (opus 4.8)	0.889	—	0.660	—	0.236	—	0.595
claude-sonnet-4-6	0.915	2.69	0.662	3.42	0.148	2.09	0.575
gpt-5.5‡	0.927	1.38	0.490	0.94	0.216	1.35	0.544
qwen3.5-397b	0.927	1.29	0.555	3.18	0.148	1.45	0.543
apodex-1.0	0.927	1.53	0.504	3.06	0.148	0.93	0.526
AlphaFold 3 [21] + symmetry expansion* §	0.927	—	0.740	—	0.148	—	0.605
Boltz-1 [92] + symmetry expansion*	0.867	—	0.695	—	0.074	—	0.545
AlphaFold 2 [93] + symmetry expansion*	0.901	—	0.732	—	0.073	—	0.569
CapBuild [94]	—	—	0.593	—	—	—	—
Template docking [95]	—	—	—	—	0.252	—	—

‡ gpt-5.5 has low HDS6 scores across all three instances, it consistently skips the self-validation step the task requires and submits the cheapest answer. Loop: folds every target with one default engine without checking or comparison. Assembly: copies the highest sequence-identity template's coordinates verbatim for every target. Complex: hardcodes outcome with severe collapses.

* Folding models cannot predict the 60-mer shell directly. For the assembly instance we fold the single subunit with the named engine and expand it onto the icosahedral operators of a fixed AAV capsid template.

§ published SOTA.

Table 19: Generative design. Designing novel AAV9 capsid peptides under a limited scoring budget, across three tissue instances (brain, spinal cord, heart) and two receptor instances (LY6A, LY6C1). A design counts as a hit if it is novel, manufacturable, and its predicted selectivity for the target ranks in the top of all sequences; the score is the fraction of the 1000 designs that are hits. All three reference methods are our own reproductions under the identical episode budget (1000 submitted designs, 100 scoring queries), differing only in the generator.

	Tissue selectivity						**Receptor specificity**				
	brain		spinalcord		heart		LY6A		LY6C1		**mean**
model	score	HDS6	score	HDS6	score	HDS6	score	HDS6	score	HDS6	
kimi-k3	0.195	3.28	**0.302**	**3.70**	0.223	3.75	**0.132**	**3.53**	0.117	**3.71**	0.194
glm-5.2	0.249	**3.73**	0.265	3.57	**0.242**	3.42	0.082	2.62	0.117	3.28	0.191
apodex-1.1	**0.259**	3.26	0.265	3.50	0.220	**3.81**	0.035	3.21	**0.122**	2.36	0.180
gpt-5.5	0.117	1.94	0.265	2.33	0.192	2.55	0.073	2.53	0.061	1.85	0.142
deepseek-v4-pro	0.112	2.91	0.257	3.03	0.204	3.06	0.038	1.61	0.032	2.58	0.129
claude-sonnet-4-6	0.078	2.95	0.201	3.02	0.108	3.00	0.041	3.35	0.043	2.55	0.094
qwen3.5-397b	0.066	2.58	0.175	1.67	0.099	2.28	0.054	2.25	0.030	2.25	0.085
Claude Code (sonnet 4-6)	0.077	—	0.187	—	0.100	—	0.035	—	0.002	—	0.080
apodex-1.0	0.025	2.37	0.228	2.94	0.062	2.78	0.039	1.67	0.008	1.92	0.072
AAVGen[22]	0.104	—	0.247	—	0.160	—	0.026	—	0.009	—	0.109
AAVDiff[23][§]	0.113	—	0.267	—	0.177	—	0.021	—	0.003	—	0.116
ALICE[19]	0.093	—	0.247	—	0.168	—	0.024	—	0.019	—	0.110

[§] published SOTA.

C. Full harness × model result matrix

For each of the 11 evaluated LLM environments, Tables 20–30 report every harness–model score underlying Section 4.3. Rows are harnesses, columns are foundation models (Opus-4.8, GPT-5.6-sol, DeepSeek-v4-pro, GLM-5.2, Kimi-k3, and Apodex-1.0 (397B)); a blank cell means the configuration was not run. Superscript letters flag a non-capability outcome (harness bug, infrastructure fault, refusal, or gate failure), explained in a note beneath the table. Claude Code and Codex CLI are run on a single model each (Opus-4.8 and GPT-5.6-sol respectively) by construction, not by omission.

Of the 11 LLM environments (Tables 20–30), seven are used for the main-text aggregates in Section 4.3. The remaining four are excluded for failing one of the pre-specified inclusion criteria: `nanogpt-speedrun` exhibits severe floor-level scores across nearly all configurations; `swe-juice` has substantial harness–model coverage gaps; `model-weakness-internal-benchmark` shows incomplete or non-compliant submissions across several harness–model cells; and `llm-determinism` is dominated by a single-model effect rather than a harness effect. All four environments are nonetheless reported in full below.

Table 20: Harness–model results on `LLM-pretrain-data-probe`, an environment that uses capture–recapture population estimation to probe the coverage and poison sources of a pretraining corpus. Each score is the mean over six episodes (arXiv / bioRxiv / medRxiv, at two coverage rungs), combining the accuracy of the population estimate with decoy-source detection, procurement outcome and budget efficiency, gated on corpus contamination.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.563	—	—	—	—	—
Codex CLI	—	0.294[a]	—	—	—	—
DeerFlow	0.491	0.601	0.447	0.454	0.351	0.395
OpenHands	0.442	0.534	0.444	0.432	0.379	0.459
A-Evolve	0.553	0.585	0.506	0.567	0.609	0.451
ApodexHarness	0.553	0.506	0.552	0.385	0.601	0.546

[a] Codex scores 0 on three episodes: its tight shell-loop probing exhausts the environment's per-episode request limiter, which also carries its model traffic. It averages 0.589 on the other three.

Table 21: Harness–model results on `nanogpt-speedrun`, an environment that trains nanoGPT to reach a target validation loss as quickly as possible. A submission is scored only if it simultaneously passes both a loss gate and a speed gate; otherwise the score is zero. Each cell is the mean over the environment's three episodes (`main`, `fw_slice_b`, and `fineweb_edu`).

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.009	—	—	—	—	—
Codex CLI	—	0.014	—	—	—	—
DeerFlow	0.062	0.000[b]	0.109	0.000[ab]	0.005	0.000[b]
OpenHands	0.123	0.147	0.005	0.056[a]	0.033	0.025[a]
A-Evolve	0.025	0.005	0.004	0.018	0.000[b]	0.037
ApodexHarness	0.016	0.000[ac]	0.000[c]	0.108	0.076	0.000[b]

[a] Mean over two episodes; the third returned no result. [b] Zero from mixed per-episode causes; where the loss gate was met, the speed gate was missed by under 13 s. [c] No episode produced a submittable training run.

Table 22: Harness–model results on `pretrain-data-filter`, an environment that trains a multi-head quality scorer to filter pretraining data. Scores combine filtering quality, measured per dimension against a reference by Pearson correlation, with a throughput gate that zeroes the score if not met, and are averaged over the environment's two canonical episodes.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.485	—	—	—	—	—
Codex CLI	—	0.497	—	—	—	—
DeerFlow	0.234[b]	0.491	0.000[a]	0.217[b]	0.416	0.000[a]
OpenHands	0.446	0.215[b]	0.000[a]	0.223[b]	0.000[c]	0.224[b]
A-Evolve	0.457	0.532	0.244[b]	0.000[a]	0.254[b]	0.000[a]
ApodexHarness	0.462	0.429	0.000[a]	0.241[b]	0.154[b]	0.354

[a] Zero in both episodes: the submission missed a hard gate — most often it was not self-contained, loading a backbone from a path the offline scoring sandbox does not mount. [b] Zero in one of the two episodes, so the mean is roughly half the working episode's score. [c] One episode only; the other returned no result (harness SDK fault) and is excluded.

Table 23: Harness–model results on `operator-align`, an environment that runs TransformerEngine on a GPU to detect misaligned or mismatched numerical operators. Scores are the pooled outcome $F_1 \times (0.5 + 0.5 \cdot \text{trigger\_rate})$ over all 33 instances: detections are pooled across instances before scoring, never averaged per instance.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.722	—	—	—	—	—
Codex CLI	—	0.491	—	—	—	—
DeerFlow	0.663	0.659	0.659	0.675	0.725	0.714
OpenHands	0.740	0.409	0.592	0.631	0.646	0.716
A-Evolve	0.722	0.662	0.628	0.712	0.720	0.641
ApodexHarness	0.727	0.734	0.657	0.653	0.743	0.607

Table 24: Harness–model results on `model-weakness-internal-benchmark`, an environment for building STEM-domain internal benchmark suites targeting model weaknesses. Submitted evaluation harnesses are scored against an internal reference suite on a 0–1 scale reflecting benchmark quality. Each entry averages scores against two held-out reference models, Qwen-3.5-35B and Qwen-3-30B.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.160	—	—	—	—	—
Codex CLI	—	0.097	—	—	—	—
DeerFlow	0.000[a]	0.138	0.088	0.085	0.000[b]	0.020
OpenHands	0.089	0.195	0.044	0.044	0.000[b]	0.057
A-Evolve	0.065	0.117	0.048	0.046	0.031	0.078
ApodexHarness	0.355	0.102	0.070	0.108	0.000[b]	0.068

[a] Timed out due to a slow multi-round loop in one of the two settings. [b] Did not follow the submission protocol.

Table 25: Harness–model results on the `rl-recipe` environment. This environment assesses submitted reinforcement-learning recipes via held-out performance improvement over a baseline recipe. Reported scores use a process-imputed proxy outcome derived from HDS6 metrics (imputation correlation with the true held-out outcome: $r = 0.979$), combined with anti-hacking penalties and cost terms, normalized to a 0–1 scale. Entries average the environment's two canonical episodes (`dapo_math`, `ifbench`); Codex CLI is `dapo_math` only, as its `ifbench` episode submitted no recipe. Since the true held-out outcome spans just 0.626–0.715, these scores are dominated by the anti-hacking term.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.751	—	—	—	—	—
Codex CLI	—	0.800	—	—	—	—
DeerFlow	0.821	0.791	0.748	0.724	0.789	0.731
OpenHands	0.747	0.797	0.776	0.805	0.732	0.732
A-Evolve	0.751	0.779	0.709	0.742	0.793	0.708
ApodexHarness	0.741	0.804	0.749	0.732	0.807	0.697

Table 26: Harness–model results on `posttrain-data-dedup`, an environment that deduplicates post-training data and removes contamination from a held-out test set. Scores combine deduplication F_1, de-contamination leak recall, and an over-removal rate. Each entry averages three instances of differing difficulty.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.386	—	—	—	—	—
Codex CLI	—	0.296	—	—	—	—
DeerFlow	0.304	0.269	0.310	0.360	0.266	0.319
OpenHands	0.233	0.272	0.249	0.273	0.309	0.273
A-Evolve	0.307	0.338	0.309	0.416	0.402	0.319
ApodexHarness	0.316	0.304	0.279	0.252	0.234	0.301

Table 27: Harness–model results on `swe-juice`, an environment that fine-tunes Qwen3-8B so its reasoning length is controllable by a reasoning-effort knob without losing agentic SWE issue-resolution ability. Scores combine effort-to-length controllability with the issue-resolution rate over hidden GitHub issues.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.000[a]	—	—	—	—	—
Codex CLI	—	0.210	—	—	—	—
DeerFlow	0.362	0.215	0.000[b]	0.462	0.399	0.000[c]
OpenHands	0.096	0.090	0.244	—[d]	—[d]	0.000
A-Evolve	0.493	—[e]	0.243	0.095	0.000	0.243
ApodexHarness	0.245	0.246	0.000[f]	0.000[g]	0.244	0.248

[a] No submission. [b] Harness bug killed training. [c] Kept exploring, training never started. [d] Model–harness compatibility fault. [e] OpenAI refused to run (safety flag). [f] ApodexHarness's history truncation drops the reasoning-content field, causing a 400 from the model API. [g] Refused on content-policy grounds.

Table 28: Harness–model results on `solution-verifier`, an environment that verifies whether a proposed code-repair patch is correct across pick (candidate selection) and rank (candidate ordering) instances drawn from OpenSWE and ACR. Scores are the binary correctness of the pick/rank verdict, averaged over all instances for every harness–model pair. Cells where the solver engaged but never submitted are scored zero, per the environment's submission rule.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.80	—	—	—	—	—
Codex CLI	—	0.90	—	—	—	—
DeerFlow	0.80	0.80	0.75	0.65	0.90	0.80
OpenHands	0.75	0.95	0.75	0.60	0.80	0.50
A-Evolve	0.85	0.90	0.75	0.80	0.75	0.50
ApodexHarness	0.85	0.95	0.70	0.85	0.90	0.60

Table 29: Harness–model results on `LLM-rollout-verify`, an environment that builds a three-way verifier for math solutions (correct / incorrect / ill-posed): the agent authors a `judge_one` function and the environment runs it over the corpus. Scores are macro-F_1 on a hidden held-out split, discounted logarithmically for output-token cost per question.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.375	—	—	—	—	—
Codex CLI	—	0.636	—	—	—	—
DeerFlow	0.424	0.452	0.347	0.477	0.531	0.312
OpenHands	0.687	0.742	0.359	0.441	0.490	0.310
A-Evolve	0.341	0.740	0.354	0.577	0.547	0.313
ApodexHarness	0.625	0.416	0.506	0.550	0.625	0.343

Table 30: Harness–model results on `llm-determinism`: an environment that finds and fixes sources of non-determinism in a pinned SGLang deployment, graded against a hidden bit-exactness suite with a no-regression check as a hard gate. Each cell is the *best-of-n* over that cell's repeats. A dash is a cell that was not run.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	GLM-5.2	Kimi-K3	Apodex-1.0
Claude Code	0.050	—	—	—	—	—
Codex CLI	—	0.480[a]	—	0.000[a]	—	—
DeerFlow	0.000[b]	0.000[b]	0.000[b]	0.000[b]	0.000[b]	0.000[b]
OpenHands	0.000[a]	0.856	0.000[b]	0.000[a]	0.000[b]	0.000[b]
A-Evolve	0.000[b]	0.100	0.067	0.000[b]	0.100	0.000[b]
ApodexHarness	0.000[a]	0.856	0.000[a]	0.820	0.856	0.100

Why a cell is zero: 8 of the 23 recorded zeros are genuine, the rest are artifacts, so an unannotated zero would not mean what it appears to. None is caused by a confirmed regression — the no-regression gate fired five times across 72 episodes and no firing was ever confirmed. [a] Genuine: the run completed and satisfied none of the required fixes. [b] Harness artifact — no scorable submission ever reached the world: early stopping or a tool-frequency guard (DeerFlow), an iteration or wall-clock cap.

Table 31: Harness–model results on `clinical-sap-tlf`, an environment that, given a statistical analysis plan and real CDISC ADaM data, reproduces the pre-specified tables, listings, and figures for a clinical trial, on its TEAE-summary, MMRM-efficacy, and time-to-event tracks. The score is the rigour of the documented statistical judgement, gated to zero unless every pre-specified number is reproduced exactly and every audit gate passes. Tracks use different statistical methods and are never averaged.

Harness (model)	teae	mmrm	tte
Claude Code (Opus-4.8)	0.000[a]	0.000[b]	0.304
Codex CLI (GPT-5.6-sol)	0.670	0.437	0.406
DeerFlow (Opus-4.8)	0.953	0.964	0.979
DeerFlow (GPT-5.6-sol)	0.982	0.622	0.911
DeerFlow (DeepSeek-v4-pro)	0.000[e]	0.000[e]	0.000[e]
DeerFlow (Kimi-k3)	0.773	0.000[b]	0.979
DeerFlow (GLM-5.2)	0.947	0.973	0.954
OpenHands (Opus-4.8)	1.000	0.676	1.000
OpenHands (GPT-5.6-sol)	1.000	0.796	0.911
OpenHands (DeepSeek-v4-pro)		—[f]	
OpenHands (Kimi-k3)	0.941	0.911	1.000
OpenHands (GLM-5.2)	0.947	0.000[c]	1.000
A-Evolve (Opus-4.8)	0.667	0.000[d]	0.656
A-Evolve (GPT-5.6-sol)	0.659	0.410	0.389
A-Evolve (DeepSeek-v4-pro)	0.000[a]	0.000[b]	0.653
A-Evolve (Kimi-k3)	0.674	0.000[a]	0.400
A-Evolve (GLM-5.2)	0.649	0.000[b]	0.000[a]

48 cells, reported after eight environment and contract fixes that removed thirteen spurious zeros. A zero is usually an audit-gate zero on a table that is numerically correct. [a] Hard-gate violation: figures reported without supporting executed code, an undeclared assumption, or submitted code that does not re-run in a clean container. [b] Five of seven numbers matched; the committed reference p-values carry an implicit Dunnett adjustment the plan never specifies, and the gate admits no partial credit — six configurations stop at exactly this signature. [c] One of seven numbers matched. [d] Ran 35 actions but never submitted. [e] Harness-side defect: its client drops the `reasoning_content` the model requires to be echoed back, killing the run at the first action; the same model completes all three tracks under A-Evolve. [f] Not runnable for the same reason, so the pair produced no episode.

Table 32: Harness–model results on `drug-repurposing-reformulation` (drr), an environment that scores a submitted promisingness and confidence estimate for a drug–disease pair against held-out clinical outcomes, evaluated by ranking and accuracy of the repurposing prediction. Each cell is the mean per-episode calibrated score over the 100 public instances, one run per instance.

Harness	Opus-4.8	GPT-5.6-sol	DeepSeek-v4-pro	Kimi-k3	GLM-5.2
Claude Code	0.942	—	—	—	—
Codex CLI	—	0.963	—	—	—
DeerFlow	0.933	0.914	0.888	0.907	0.899
OpenHands	0.865	0.926	0.916	0.956	0.934
A-Evolve	0.970	0.884	0.914	0.912	0.948

1700 episodes. Balanced-grid mean 0.9178; spread across harnesses 0.0176, across models 0.0212. Cells include submission-failure zeros (per 100 instances, in column order: DeerFlow 2/5/5/4/4, OpenHands 10/4/3/0/2, A-Evolve 0/8/3/4/0; Claude Code 2, Codex 0). Excluding those zeros the harness spread collapses to 0.0098 and the ordering flips, so this environment separates submission reliability rather than judgement quality — GPT has the highest judgement quality (0.9625 excluding zeros) but the most submission failures. Instance difficulty dominates: the between-instance standard deviation (0.133) is about four times the between-cell one (0.034). This is the per-episode calibrated score, not the official drep normalized/combined metric.

Table 33: Harness–model results on AAV capsid enrichment, an environment that trains a sequence-to-enrichment model for AAV9 588-loop 7-mer variants, on three generalisation tracks: mouse multi-organ tropism, cross-species transfer to macaque liver, and human liver-cell lines *in vitro*. Scores are the per-assay Pearson correlation between predicted and true enrichment, aggregated into a track total; tracks are reported separately and never averaged.

Harness (model)	mouse	macaque	in vitro
Claude Code (Opus-4.8)	0.680	0.577	0.756
Codex CLI (GPT-5.6-sol)	0.583	0.523	0.754
DeerFlow (Opus-4.8)	0.540	0.485	0.731
DeerFlow (GPT-5.6-sol)	0.602	0.580	0.742
DeerFlow (DeepSeek-v4-pro)	0.549	0.501	0.000[a]
DeerFlow (Kimi-k3)	0.698	0.561	0.741
DeerFlow (GLM-5.2)	0.633	0.320	0.727
OpenHands (Opus-4.8)	0.575	0.617	0.748
OpenHands (GPT-5.6-sol)	0.409	0.631	0.762
OpenHands (DeepSeek-v4-pro)	0.502	0.585	0.707
OpenHands (Kimi-k3)	0.515	0.494	0.750
OpenHands (GLM-5.2)	0.517	0.589	0.747
A-Evolve (Opus-4.8)		—[b]	
A-Evolve (GPT-5.6-sol)	0.628	0.472	0.749
A-Evolve (DeepSeek-v4-pro)	0.371	0.594	0.649
A-Evolve (Kimi-k3)	0.468	0.568	0.751
A-Evolve (GLM-5.2)	0.516	0.633	0.612

Reference points per track: k-mer lookup floor 0.182 / 0.052 / 0.000; Fit4Function paper reproduction 0.532 / 0.614 / 0.719. Every scored cell clears the k-mer floor and most reach or exceed the paper reproduction; the cross-species macaque track is the discriminating one. GPT's first-pass timeouts on this host-bash world were recovered by a low-concurrency re-run, confirming a submission-discipline rather than a capability limit. [a] Grounding failure: the agent stopped to ask the user instead of acting, and stronger prompting did not recover it — the only unrecovered cell in the sweep. [b] A-Evolve's provider layer rejects non-OpenAI model names, so all three cells are unavailable by construction rather than failures.

Table 34: Harness–model results on AAV capsid structure, an environment that predicts three-dimensional structure from sequence for the VR-loop monomer, the full 60-mer capsid assembly, and the VP–ligand complex instances of the task. Scores compare the predicted structure against a time-cut-off held-out reference; the complex instance is ungated.

Harness (model)	loop	assembly	complex
Claude Code (Opus-4.8)	0.929	0.692	0.136
Codex CLI (GPT-5.6-sol)	0.927	0.774	0.379
DeerFlow (Opus-4.8)	0.927	0.000[a]	0.163
DeerFlow (GPT-5.6-sol)	0.927	0.664	0.163
DeerFlow (DeepSeek-v4-pro)	0.927	0.000[b]	0.000[c]
DeerFlow (Kimi-k3)	0.927	0.000[d]	0.000[c]
DeerFlow (GLM-5.2)	0.928[e]	0.000[d]	0.163
OpenHands (Opus-4.8)	0.927	0.668	0.120
OpenHands (GPT-5.6-sol)	0.930	0.819	0.342
OpenHands (DeepSeek-v4-pro)	0.927	0.000[b]	0.163
OpenHands (Kimi-k3)	0.925	0.000[d]	0.017
OpenHands (GLM-5.2)	0.927	0.709	0.163
A-Evolve (Opus-4.8)		—[f]	
A-Evolve (GPT-5.6-sol)	0.930	0.591	0.128
A-Evolve (DeepSeek-v4-pro)	0.927	0.704	0.136
A-Evolve (Kimi-k3)	0.927	0.000[d]	0.365
A-Evolve (GLM-5.2)	0.927	0.634	0.163

Metrics: loop = VR-loop lDDT on the low-homology (<80%) stratum, gated on beating the nearest-template baseline (lDDT 0.709) with coverage ≥ 0.9; assembly = median$(\exp(-\text{RMSD}/5\,\text{Å}) \times \text{stability})$ under clash gating, scaled by the fraction of the 60-mer actually built; complex = mean over an antibody (8 targets) and a receptor (7 targets) probe of $0.5 \cdot \text{DockQ} + 0.5 \cdot$ interface Fnat, with no pass/fail gate. Each cell is the best of up to three attempts; no leakage fingerprint triggered in any of the 79 episodes. The loop track is saturated — 11 of the 16 scored cells sit at 0.927 and the submitted structures are byte-identical across harnesses, because the default fold path is deterministic and cached; six cells likewise sit at exactly 0.163 on complex, the floor for submitting the default fold unchanged. Assembly is the only strongly discriminating axis in this round. [a] The scorer found no submission to grade. [b] All 60 chains were placed but whole-capsid RMSD is 100–155 Å, so the RMSD score is zero. [c] All 15 complexes were graded but DockQ and interface Fnat are both zero — verified to be a real failure, not a silently swallowed export error. [d] None of the 60-mer was built. [e] Score is fine but only 5 of 12 targets were folded (coverage 0.417), so the correctness gate fails. [f] A-Evolve's provider layer rejects non-OpenAI model names.

Table 35: Harness–model results on AAV sequence design, an environment that designs AAV9 588-loop 7-mer peptides optimized for targets such as receptor specificity or tissue tropism, for the brain, heart, spinal cord, Ly6a, and Ly6c1 targets. Tropism targets are scored as the mean of the pass rates at the per-organ top-5%, top-2%, and top-1% selectivity bars, combined with a viability gate; receptor targets as the specific pass rate. Opus refuses the task on Claude Code and DeerFlow.

Harness (model)	brain	heart	spinalcord	ly6a	ly6c1
Claude Code (Opus-4.8)	REFUSE (all targets)				
Codex CLI (GPT-5.6-sol)	0.119	0.290	0.415	0.070	0.135
DeerFlow (Opus-4.8)	REFUSE (all targets)				
DeerFlow (GPT-5.6-sol)	0.059[a]	0.076[a]	0.221	0.062	0.000[b]
DeerFlow (DeepSeek-v4-pro)	0.025	0.051	0.138	0.010	0.013
DeerFlow (Kimi-k3)	0.040	0.102	0.241	0.021	0.015
DeerFlow (GLM-5.2)	0.090[a]	0.120[a]	0.220	0.088	0.056
OpenHands (Opus-4.8)			—[c]		
OpenHands (GPT-5.6-sol)	0.185	0.153	0.392	0.132	0.175
OpenHands (DeepSeek-v4-pro)	0.075	0.150	0.269	0.080	0.021
OpenHands (Kimi-k3)	0.100	0.095	0.214	0.076	0.038
OpenHands (GLM-5.2)	0.109	0.193	0.357	0.066	0.001
A-Evolve (GPT-5.6-sol)	0.122	0.174	0.349	0.141	0.088
A-Evolve (DeepSeek-v4-pro)	0.060	0.091	0.174	0.025	0.073
A-Evolve (Kimi-k3)	0.047	0.075	0.148	0.066	0.051
A-Evolve (GLM-5.2)	0.136	0.151	0.091	0.087	0.059

Each tropism column is the mean of that organ's two evaluation splits (same metric and scale); 84 Track-A cells in total, best attempt per cell. Generative reference baselines under the same three-bar rule: aavdiff 0.113 / 0.177 / 0.267, aavgen 0.104 / 0.160 / 0.247 and alice 0.093 / 0.168 / 0.247 for brain / heart / spinal cord; only 4 of 14 evaluated configurations beat aavdiff on the six-instance mean. The strictest bar is counting-noise dominated (median 24 qualifying designs at the 2% bar and 2 at the 1% bar out of a 1000-design budget), which is why the reported score averages the three bars. A-Evolve cannot be run with Opus (its provider layer rejects non-OpenAI model names). [a] Average includes one split that scored 0.000 because nothing was submitted. [b] Isolated incident, not reproduced on other targets under the same configuration. [c] Opus was not part of the sweep these three-bar scores were computed from; the earlier Opus figures were obtained under the superseded single top-2% bar and are not comparable.

Table 36: Harness–model results on AAV viability, an environment that predicts from protein-sequence features whether an AAV capsid variant is viable, meaning it can assemble and package its genome, on two orthogonal difficulty axes: dataset (Bryant 2021 substitutions vs. Ogden 2019 insertions and deletions) and validation-set visibility (labelled, or a hidden oracle reachable only through a metered query). Scores are prediction accuracy against hidden labels; the Bryant settings are saturated near ceiling, with limited discrimination among configurations.

	Bryant		Ogden	
Harness (model)	labelled	oracle	labelled	oracle
Claude Code (Opus-4.8)	0.947	0.968	ref[a]	ref[a]
Codex CLI (GPT-5.6-sol)	0.980	0.955	0.928	0.903
DeerFlow (Opus-4.8)	0.960	0.952	0.908	0.862
DeerFlow (GPT-5.6-sol)	0.976	0.948	0.936	0.854
OpenHands (Opus-4.8)	0.982	0.957	0.867	0.851
OpenHands (GPT-5.6-sol)	0.982	0.943	0.940	0.913
A-Evolve (Opus-4.8)		—[c]		
A-Evolve (GPT-5.6-sol)	0.986	0.970	0.930	ref[b]

Metric: mean per-distance AUROC, with the passing threshold set by the environment's additive baseline (Bryant 0.885, Ogden 0.851; literature reference cap-PLM/ESM2 $\approx$ 0.897 on Bryant). All 33 scored cells pass and the Bryant settings broadly exceed the literature reference; the difficulty gradient Ogden $>$ Bryant and oracle $>$ labelled is visible in both scores and pass rates. Eleven first-pass zeros were concurrency artefacts and recovered on a low-concurrency re-run; capability failures: zero. Effort is uncorrelated with score (10 to 1970 actions). DeepSeek-v4-pro, Kimi-k3 and GLM-5.2 were not run on this environment. REF = deterministic provider-side safety refusal on the Ogden capsid-engineering framing, harness-specific and not a capability result. [a] Anthropic usage-policy refusal (one action, never submitted). [b] OpenAI `invalid_prompt`, reproduced at three concurrency levels. [c] A-Evolve's provider layer rejects non-OpenAI model names.

D. Drug Repurposing and Reformulation Trace Example

This appendix reports one evaluation episode end to end: the full recorded reasoning trace, and the full process score with every sub-rubric shown.

Table 37: Episode summary. The reference column is never visible to the judging panel.

Field	Value
Drug–disease pair	Dostarlimab → Stage II Breast Cancer
System score	**0.3**
System confidence	0.48
Held-out reference	**0.15** (lower bound; see below)
Reference clinical phase	Phase 2
Reference outcome	Ongoing
Benchmark score	**1.0** (error **0.0**)
Integrity gate	**PASS**
Recorded steps	18 (9 reasoning, 8 tool calls, 1 submission)
Key claims submitted	7

D.1. The task as given

Before any reasoning, this is the complete brief the system received — the enforced offline-evaluation policy, the analyst role, the required workflow, the evidence-grounding rules, and the output contract it had to satisfy. It is reproduced verbatim.

TASK INPUT • BRIEF GIVEN TO THE SYSTEM

OFFLINE EVALUATION MODE (ENFORCED)

Do not use the open internet to look up this pair's recent outcome: no general web search, no fetching arbitrary URLs, no browsing news, and no directly hitting public services (e.g. ClinicalTrials.gov, PubMed, preprint servers) for the trial's result. The held-out outcome is recent, so retrieving it that way would invalidate the evaluation; those paths are blocked.

This does NOT restrict the sanctioned environment. Only capabilities listed by the selected bundle manifest and bound by the active access adapter are allowed. That includes tool_A's Stage-1 `provider_A` (live path-finding over the structured bio-graph) and any resident inference-service interfaces explicitly exposed by that bundle; copied documentation alone never grants another capability.

You are a senior drug-repurposing analyst acting as an autonomous research agent.

You assess drug-repurposing candidates. For each drug-disease pair you are asked about, you estimate how far that pair is likely to progress along the FDA clinical-approval spectrum, and how confident you are in that estimate.

Treat this as a genuine discovery task. Start with one bounded tool_A Stage-1 call using `--tier extra --output <file> --summary`; its summary is the canonical and sufficient decision surface. Do not inspect the artifact/filesystem or dump `coarse.provider_results/fused_graph`. Run a supplied `detail_command` once only for a load-bearing page marked `truncated_for_summary: true`. An absent page is an evidence limitation, not a missing field.

A node budget is each provider's frontier-expansion budget, not its returned-node count. A `provider_A partial` is usable bounded coverage---not a provider failure or negative evidence---when `error` is null, `timed_out` is false, and `budget_exhausted` is true. Answer when resolved anchors have grounded connections. Otherwise, decision-critical coverage permits one next-tier rerun into a new file; stable IDs or budget exhaustion alone do not. Stage 2 is optional: use one suggested `tool_A expand` route only if it could change the score or confidence, never by rote or through bulk `--stage2 auto`. Never repeat a tier, widen an error/timeout, rebuild/install providers, or use `bash -lc`, pipelines, or retry loops. Preserve conflicts and operational limitations in your conclusion.

Check the operating manual on how to interact with the environment below (in the <manual_B> section).

Definition of required scores for output

PROMISINGNESS (a number in [0, 1]) is the pair's predicted position on the clinical-approval spectrum. It is scored against the pair's real, held-out clinical outcome and is also used to rank candidates, so calibrate it to reflect true relative promise rather than enthusiasm.

CONFIDENCE (a number in [0, 1]) is your estimate of prediction quality: how close your promisingness score is to the true held-out value. It is scored as 1 - (confidence - quality) ^ 2, where quality = 1 - prediction_error ^ 2. The optimal strategy is to set confidence equal to your expected prediction quality. Report low confidence when evidence is thin or conflicting.

Anchor your promisingness estimate to the tier scale:

Reference points (how the held-out outcome maps to the target score): 0.00 preclinical only (or failed@Ph1) 0.15 passed Phase 1 (safety established) 0.45 passed Phase 2 (efficacy signal) 0.65 passed Phase 3 (pivotal) or off-label established 1.00 FDA approved on-label A Phase-3 success (∼0.65) is the strongest single predictor of approval; a full approval scores 1.00. Interpolate between points.

Scoring rules:

- A successful but not-yet-approved Phase-1, Phase-2, or Phase-3 result is a LOWER BOUND on eventual program potential, not a terminal point target. Predicting below the phase actually passed is penalised; predicting above it is fine when the evidence supports further progression.
- Genuinely terminal outcomes use an exact target. Failing at a later phase caps the credit at the last passed phase --- e.g. a Phase-3 failure scores exactly 0.45 (Phase 2 was passed), not 0.65.
- Ongoing trials should score at LEAST the tier of the last phase they passed. A drug currently in Phase 3 has already passed Phase 2, so its score should be >= 0.45. Predicting below that threshold is penalised; predicting above it is fine (you are estimating its eventual outcome).
- Interpolate between tier points when the evidence suggests an intermediate position.

Evidence grounding (this is graded --- cite like a RAG system):

- Ground every load-bearing claim in a SPECIFIC source you actually retrieved from the environment --- a PubMed id (PMID: ########), a trial registry id (NCT#########), a knowledge-graph edge/id, a database record (e.g. ChEMBL / Open Targets / drug label), or the exact tool + query you ran. Prefer sources you pulled from the tools over parametric memory.
- For each such claim, keep the SUPPORTING SNIPPET --- the specific retrieved finding (a quoted line, a value, an edge) that backs it --- so the claim is self-justifying and independently checkable.
- You will report these as a structured `key_claims` list in the output contract (claim + source + support). A downstream verifier checks whether each claim's cited source actually supports it, so uncited or unsupported claims lower the score --- cite precisely, quote faithfully, do not invent sources.
- If a claim rests only on model knowledge with no retrieved source, set its source to "parametric", leave `support` empty, and lower your confidence.

TASK: Assess the drug-repurposing potential of the following drug-disease pair.

drug: Dostarlimab disease: Stage II Breast Cancer drug_id: <drug_id> disease_id: <disease_id>

Investigate the pair carefully, using the available tools if helpful. Write a comprehensive report, then provide a promisingness score, a confidence score, and a concise rationale, following the output contract exactly.

Output contract

Write your full, comprehensive report first. Then, as the VERY LAST thing in your response, emit exactly ONE fenced code block tagged `dreg-result` containing a single JSON object with exactly these keys:

```
{
  "promisingness_score": <number between 0 and 1>,
  "confidence_score": <number between 0 and 1>,
  "rationale": "<concise justification: key evidence, mechanism, and what drove the score; 3-8
sentences>",
  "key_claims": [
    {
      "claim": "<one load-bearing factual claim that drove your score>",
      "source": "<the specific retrieved source: PMID:######## | NCT######### | a KG edge/id | a DB
record e.g. ChEMBL:... / OpenTargets:... / drug label | tool:<name>+<query> | 'parametric' if only
model knowledge>",
      "support": "<the exact retrieved finding/quote that backs this claim; empty string if source is
parametric>"
    }
  ]
}
```

Rules:

- The block MUST be valid JSON (double-quoted keys and string values, no trailing commas, no comments). Escape any

quotes/newlines inside strings.

- `promisingness_score` and `confidence_score` are plain numbers in [0, 1].
- `key_claims`: 3-8 entries covering the MAIN drivers of your score. Each claim must carry the specific `source` you retrieved it from and the `support` snippet that backs it (see "Evidence grounding" above). A downstream verifier scores whether each claim's cited source supports it --- so cite precisely, quote faithfully, and never invent a source. Use "parametric" only when you truly have no retrieved evidence, and lower `confidence_score` accordingly.
- Emit the block only once, at the end. Do not wrap it in extra prose after it.

the execution environment transport adapter

This evaluation uses the execution environment actions. Do not finish with only the fenced JSON described above. Submit the same canonical JSON object with `act submit --params '{"submission": {...}}'`. Use `act domain-tool` for the bounded the domain toolkit tools documented in <toolkit_dir>.

TASK INPUT • REQUIRED OUTPUT SCHEMA

```
{
  "promisingness_score": "number in [0,1]",
  "confidence_score": "number in [0,1]",
  "rationale": "string",
  "key_claims": "list of {claim, source, support}"
}
```

TASK INPUT • EVALUATION CONTRACT

```
the domain toolkit-eval/v1
```

TASK INPUT • SCORING SCALE

```
{
  "name": "approval-spectrum-v2",
  "values": [
    0.0,
    0.15,
    0.45,
    0.65,
    1.0
  ]
}
```

D.2. Full reasoning trace

Every recorded step is shown. Reasoning turns are grey, tool calls blue, the failed call red, and the submission gold. Tool calls are shown verbatim, with the returned payload trimmed only for page width. Model reasoning turns are listed but their content is withheld: the judging panel never sees the system's private chain of thought, only the actions and outputs it produced.

REASONING • STEP 1 • Orient on the episode brief and output contract

ACTION EMITTED

```
read {"path":"<manual_C>"}
```

REASONING • STEP 2 • Decide to read the manuals before retrieving anything

ACTION EMITTED

```
bash {"command":"act inspect-task","timeout":120}
```

TOOL CALL • STEP 3 • Inspect the canonical task record

RETURNED

The `description` field is the task brief reproduced in Section D.1 above and is omitted here to avoid duplication; every other field is shown.

RETURNED

```
{
  "task": "pointwise-promisingness",
  "drug": "Dostarlimab",
  "disease": "Stage II Breast Cancer",
  "drug_id": "<drug_id>",
  "disease_id": "<disease_id>",
  "submission_schema": {
    "promisingness_score": "number in [0,1]",
    "confidence_score": "number in [0,1]",
    "rationale": "string",
    "key_claims": "list of {claim, source, support}"
  },
  "tier_scale": {
    "name": "approval-spectrum-v2",
    "values": [
      0.0,
      0.15,
      0.45,
      0.65,
      1.0
    ]
  },
  "eval_contract": "the domain toolkit-eval/v1",
  "workdir": "<workdir>",
  "staged_files": [
    "instance.json",
    "toolkit/"
  ],
  "toolkit": {
    "access_adapter": "<manual_A>",
    "operating_manual": "<manual_B>",
    "skills_dir": "<skills_dir>",
    "manifest": "<manifest>"
  },
  "note": "Bootstrap complete. Next read toolkit.access_adapter, then toolkit.operating_manual;
follow them and submit the canonical the domain toolkit fields."
}
```


REASONING • STEP 4 • Register the enforced offline-evaluation policy

ACTION EMITTED

```
bash {"command":"act bash --cmd 'cat <manual_A>'","timeout":120}
bash {"command":"act bash --cmd 'cat <manual_B>'","timeout":120}
```

TOOL CALL • STEP 5 • Read the environment access manual

CALL

```
cat <manual_A>
```

RETURNED

exit_code: 0 stdout:

Access adapter --- the execution environment (isolated SUT)

This file owns only the execution environment execution mechanics. ../<manual_B> owns the the domain toolkit retrieval and inference policy.

Actions

Every action is metered and recorded:

- `act inspect-task` returns the canonical task, visible instance, and output schema.
- `act bash --cmd '<command>'` runs notes or small computations in the isolated the sandbox.
- `act domain-tool --params '<json>'` invokes a structured, allowlisted the domain toolkit capability on the trusted tool host.
- `act submit --params '{"submission": {...}}'` records the final answer and ends the episode.

the sandbox and paths

- `bash` works only under `<workdir>`. The supplied files are `<instance_file>` and `<toolkit_dir>`; files you create are ephemeral.
- The the sandbox has no network, host mounts, the domain toolkit checkout, benchmark data, hidden reference, host tool workspace, Docker socket, or host GPU devices.
- Never use a host session, mounted-data, checkout, or the execution environment episode path.
- Persist the evaluated answer with `submit`, not a local file.

the domain toolkit tool boundary

tool_A requests have this envelope:

```
act domain-tool --params \
  '{"tool":"tool_A","command":"tool_A ..."}'
```

inference-service requests have this envelope:

```
act domain-tool --params \
  '{"tool":"tool_B","operation":"nim","arguments":{"method":"GET","port":8000,"path":"/openapi.json"}}
'
```

- Tool output paths must be virtual paths below `<artifact>`. The adapter maps them to this episode's private host-side workspace.
- `bash` cannot read tool artifacts. Use the returned summary or a policy-permitted inspection request through `domain-tool`.
- Only capabilities in `<manifest>` are callable.
- Requests are parsed into fixed arguments. No shell, SSH, pipelines, command substitution, arbitrary executable, URL, or filesystem path reaches the tool host. inference-service calls are limited to the configured local port range.

{ "stderr": "" }

TOOL CALL • STEP 6 • Read the domain toolkit manual

CALL

```
cat <manual_B>
```

RETURNED

exit_code: 0 stdout:

the domain toolkit biomedical toolkit in the execution environment

This file owns only the domain toolkit's retrieval and inference workflow. The the execution environment bootstrap brief owns episode order; [`access/<manual_A>`](access/<manual_A>) owns action syntax, the sandbox isolation, virtual artifact mapping, and submission mechanics. No skill document is required for this focused surface.

Available tools

The catalog below is generated from the domain toolkit's benchmark tool registry. Ground-truth and verifier tools are never included.

tool_A --- Coarse-to-fine biomedical pathfinding across broad KGs and typed specialist providers.

tool_B --- Vendored NVIDIA tool_B --- inference-service skills, open models, GPU libraries.

Required tool_A workflow

tool_A is the single coarse-to-fine entry point for connecting drugs, diseases, genes, proteins, pathways, and mechanisms.

Start with one bounded Stage-1 search per pair:

```
act domain-tool --params '{
  "tool":  "tool_A",
  "command":  "tool_A connect "<drug>" "<disease>" --tier extra --timeout 240 --output <artifact_1> --
summary"
}'
```

The returned `--summary` projection is the canonical decision surface. Inspect `resolved_entities`, `providers_run`, `candidate_paths`, `connections`, `load_bearing_evidence`, `relationship_pages`, `expansion_candidates`, `stage2`, and `report` there. Do not request or dump `coarse.provider_results` or `fused_graph`.

Only when a load-bearing relationship page is marked `truncated_for_summary: true`, send its supplied `detail_command` through `domain-tool` once. That command must use an existing virtual artifact:

```
act domain-tool --params '{
  "tool": "tool_A",
  "command": "tool_A inspect <artifact_1> /json/pointer"
}'
```

A provider's node budget is its frontier-expansion budget, not its returned-node count. A `provider_A` `partial` result is usable bounded coverage when `error` is null, `timed_out` is false, and `budget_exhausted` is true. It is not negative evidence.

If correctly resolved anchors already have grounded connections, answer from the summary. Otherwise, decision-critical coverage permits one next-tier rerun to a new `<artifact>` path. Never repeat a tier or widen a timed-out/error run.

Stage 2 is optional. Use one route from `expansion_candidates[].compatible_routes` only when the extra evidence could change the conclusion or confidence:

```
act domain-tool --params '{
  "tool": "tool_A",
  "command": "tool_A expand "<typed A>" "<typed B>" --provider <provider_id> --algorithm <algorithm_
id> --a-type <type> --b-type <type> --timeout 120 --output <artifact_5> --summary"
}'
```

Do not invent a route when none is suggested. Do not use bulk `--stage2 auto`, shell wrappers, pipelines, retry loops, provider installation, or provider build commands during an evaluation.

Optional tool_B inference-service inference

The the execution environment contract supports structured GET/POST inference against these resident tool-host endpoints:

| inference-service | Port | Inference path | |---|---:|---| | <model_endpoint> | 8000 | `POST <endpoint_path><model_endpoint>/predict` | | <model_endpoint> | 8001 | `POST <endpoint_path><model_endpoint>/generate` | | <model_endpoint> | 8002 | `POST /molecular-docking/<model_endpoint>/generate` | | <model_endpoint> | 8003 | `POST <endpoint_path><model_endpoint>/predict` | | <model_endpoint> | 8004 | `POST /generate` | | <model_endpoint> | 8005 | `POST /generate` | | <model_endpoint> | 8006 | `POST <endpoint_path>` | | <model_endpoint> | 8007 | `POST <endpoint_path><model_endpoint>/predict-structure-from-msa-and-template` | | <model_endpoint> | 8008 | `POST <endpoint_path><model_endpoint>/predict` | | <model_endpoint> | 8009 | `POST <endpoint_path><model_endpoint>/msa-search/predict` |

Call a inference-service through the structured adapter:

```
act domain-tool --params '{
  "tool": "tool_B",
  "operation": "nim",
  "arguments": {
    "method": "POST",
    "port": 8003,
    "path": "<endpoint_path><model_endpoint>/predict",
    "json": {}
  }
}'
```

The port must be 8000–8009, the path must be local and absolute, and the body must be JSON. If the exact deployed request schema is needed, send a structured `GET` to `/openapi.json` on that inference-service's assigned port. Do not run inference-service setup, `docker login`, `docker run`, readiness loops, or direct `curl`; deployment health is controlled by the operator.

{ "stderr": "" }

REASONING • STEP 7 • Form the retrieval plan

MODEL NARRATION

I have the operating rules. I'm going to run the required bounded tool_A Stage-1 search and base the decision on its summary unless it exposes a specific truncated page that needs one permitted detail lookup.

TOOL CALL • STEP 8 • Bounded pairwise search for the assigned pair

CALL

```
tool_A connect "Dostarlimab" "Stage II Breast Cancer" --tier extra --timeout 240 --output <artifact_1>
--summary
```

RETURNED

exit_code: 0 tool: tool_A contract_version: the domain toolkit-tools/v1 stdout: {"output_path": "<artifact_1>", "summary": {"schema_version": "tool_A/v2", "stage": "stage1", "workflow": {"review_before_next_call": true, "stage1_follow_up": "Optionally query one decision-relevant related_page_hint for more Stage-1 literature context.", "stage1_follow_up_command_template": "tool_A connect <hint.title> --a-id <hint.uid> --stage2 suggest --output <new-stage1-path> --summary", "stage2": "Only after reviewing Stage 1, optionally run one advertised ready expansion route when a deeper connection could change the conclusion."}, "artifact_inspection": {"default": "not_needed", "reason": "This projection already contains provider status, selected paths, load-bearing evidence, anchor and relationship pages, related-page hints, conflicts, and optional expansion candidates.", "exception": "Run only an item's supplied detail_command, once, when its truncated_for_summary flag is true and the omitted text is load-bearing. An absent relationship page is not a missing field and never authorizes artifact or filesystem search.", "provider_results_shape": "coarse.provider_results is an object keyed by canonical provider ID, not a list."}, "query": {"id": "query-<record>", "text": null, "entities": ["Dostarlimab", "Stage II Breast Cancer"], "input_entities": [{"id": "Dostarlimab", "name": "Dostarlimab", "type": "unknown", "species": null, "aliases": [], "resolution": {"source": "caller", "structured_input": true}}, {"id": "Stage II Breast Cancer", "name": "Stage II Breast Cancer", "type": "unknown", "species": null, "aliases": [], "resolution": {"source": "caller", "structured_input": true}}], "operation": "connect", "anchor_quality": {"trusted": false, "unresolved": [{"id": "<local_id>", "name": "stage ii breast cancer", "type": "unknown", "species": null, "aliases": ["Stage II Breast Cancer"], "resolution": {"source": "surface_form", "ambiguous": true, "providers": ["provider_A"], "provider_ids": {"provider_A": "stage ii breast cancer"}, "is_anchor": true, "input_aliases": ["Stage II Breast Cancer"]}}]}}, "resolved_entities": [{"id": "<record_1>", "name": "DOSTARLIMAB", "type": "drug", "species": null, "aliases": ["dostarlimab", "Dostarlimab"], "resolution": {"source": "provider_canonical_id", "ambiguous": false, "providers": ["provider_A", "provider_B"], "provider_ids": {"provider_A": "DOSTARLIMAB", "provider_B": "<record_1>"}, "is_anchor": true, "previous_ids": ["<local_id>"], "input_aliases": ["Dostarlimab"]}}, {"id": "<local_id>", "name": "stage ii breast cancer", "type": "unknown", "species": null, "aliases": ["Stage II Breast Cancer"], "resolution": {"source": "surface_form", "ambiguous": true, "providers": ["provider_A"], "provider_ids": {"provider_A": "stage ii breast cancer"}, "is_anchor": true, "input_aliases": ["Stage II Breast Cancer"]}}], "providers_run": [{"provider_id": "provider_A", "status": "partial", "algorithm_status": "partial", "algorithm_id": "weighted_best_first", "elapsed_ms": 71465.406, "timed_out": false, "budget_exhausted": true, "expansions": 150, "frontier_remaining": 3197, "early_stopped_on_connection": false, "effective_limits": {"max_expansions": 150, "neighbor_limit": 5, "max_depth": 5, "max_hops": 7, "top_k": 5, "timeout_seconds": 240}, "warnings": ["algorithm expansion budget exhausted after 150 frontier expansions; unvisited frontier remains"], "error": null}, {"provider_id": "provider_B", "status": "partial", "algorithm_status": "partial", "algorithm_id": "weighted_best_first", "elapsed_ms": 11516.748, "timed_out": false, "budget_exhausted": true, "expansions": 150, "frontier_remaining": 739, "early_stopped_on_connection": false, "effective_limits": {"max_expansions": 150, "neighbor_limit": 5, "max_depth": 5, "max_hops": 7, "top_k": 5, "timeout_seconds": 240}, "warnings": ["algorithm expansion budget exhausted after 150 frontier expansions; unvisited frontier remains"], "error": null}], "projection_counts": {"candidate_paths": {"shown": 0, "total": 0}, "connections": {"shown": 0, "total": 0}, "load_bearing_evidence": {"shown": 0, "total_selected_edge_ids": 0}, "relationship_pages": {"shown": 0, "total": 0, "direct_for_anchor_pair_available": false}, "anchor_pages": {"shown": 1, "resolved_anchors": 1}, "related_page_hints": {"shown": 7, "total": 21}, "expansion_candidates": {"shown": 0, "total": 0}}, "candidate_paths": [], "expansion_candidates": [], "connections": [], "load_bearing_evidence": [], "anchor_pages": [{"uid": "<record_1>", "title": "dostarlimab", "type": "drug", "path": "drugs/dostarlimab.md", "n_papers": 238, "pmids": ["39282229", "41904276", "35892341", "36005013", "0", "41811823", "36762991", "38365286", "38605944"], "text": "--- uid: <record_1> type: drug name: "dostarlimab" aliases: ["Dostarlimab", "Dostarlimab (Jemperli)"] n_papers: 238 n_claims: 388 n_claims_used: 13 grounded: false synth_model: "glm-5.2" prompt_template_version: "entity_page_v1" updated: 2026-07-06 tags: ["drug"]

dostarlimab

Overview

Dostarlimab is an FDA-approved anti-PD-1/PD-L1 antibody widely used in cancer treatment [ref-9], [ref-11]. It is an IgG4 isotype designed to avoid tumor-reactive T cell depletion, with little to no binding to Fc or complement protein C1q [ref-7]. Its pharmacokinetic profile is characterized by a 2-compartment model with time-dependent linear elimination, a mean terminal elimination half-life of 25.4 days, a mean clearance of 0.007 L/h, and a steady state volume of distribution of 5.3 L [ref-1], [ref-2]. As a tumor-agnostic therapy based on genetic changes, dostarlimab is highly relevant for drug repurposing across multiple histologies [ref-10].

Repurposing directions

Strongly supported directions

Dostarlimab is strongly supported as a treatment for [[dostarlimab → cancer]], with high-quality clinical trial and review evidence plus in vivo animal data indicating beneficial", "truncated_for_summary": true, "full_text_artifact_json_pointer": "/coarse/provider_results/provider_B/pages/0/text", "detail_command": "tool_A inspect <artifact_1> /coarse/provider_results/provider_B/pages/0/text"}], "relationship_pages": [], "related_page_hints": [{"uid": "<record>", "title": "dostarlimab → endometrial cancer", "path": "hypotheses/dostarlimab → endometrial cancer.md", "n_papers": 29, "type": "hypothesis", "

follow_up_stage": "stage1"}, {"uid": "<record_2>", "title": "PD-1", "path": "targets/PD-1.md", "n_papers": 12098, "type": "target", "follow_up_stage": "stage1"}, {"uid": "<record_1>", "title": "FC", "path": "drugs/FC.md", "n_papers": 20, "type": "drug", "follow_up_stage": "stage1"}, {"uid": "<record>", "title": "dostarlimab → locally advanced rectal cancer", "path": "hypotheses/dostarlimab → locally advanced rectal cancer.md", "n_papers": 12, "type": "hypothesis", "follow_up_stage": "stage1"}, {"uid": "<record_2>", "title": "PD-L1", "path": "targets/PD-L1.md", "n_papers": 10947, "type": "target", "follow_up_stage": "stage1"}, {"uid": "<record>", "title": "dostarlimab → rectal cancer", "path": "hypotheses/dostarlimab → rectal cancer.md", "n_papers": 11, "type": "hypothesis", "follow_up_stage": "stage1"}, {"uid": "<record_2>", "title": "programmed cell death protein 1 (PD-1)", "path": "targets/programmed cell death protein 1 (PD-1).md", "n_papers": 195, "type": "target", "follow_up_stage": "stage1"}], "stage2": {"policy": "suggest", "optional": true, "suggested_expansions": 0, "requested_routes": 0, "guidance": "Stage 1 is complete. The calling agent may answer now or request one or more compatible expansion routes only when a load-bearing weak, missing, or conflicting link could change the conclusion."}, "report": {"summary": "No complete connection path was reconstructed for Dostarlimab ↔ Stage II Breast Cancer. Provider results and unresolved gaps are retained for diagnosis.", "support_grade": "incomplete", "conflicts": [], "conflict_count": 0, "unresolved_gaps": [{"gap_id": null, "candidate_path_id": null, "subject": "<local_id>", "object": null, "required_provider": null, "reason": "unresolved_or_local_query_anchor"}], "unresolved_gap_count": 1, "optional_expansions": {"policy": "suggest", "suggested": 0, "requested": 0}, "diagnostics": {"non_load_bearing_conflict_count": 245}, "interpretation": "Provider-native scores are shown separately and were not averaged. A support grade describes evidence structure, not treatment efficacy."}, "graph_warnings": []}}

{ "stderr": "" }

REASONING • STEP 9 • Read the negative result; note the unresolved disease anchor

REASONING SUMMARY

Deciding truncated load-bearing detail inclusion

MODEL NARRATION

Stage-1 found no direct reconstructed connection, but the drug anchor page is truncated and is the only available support for mechanism/status, so I'm using the one permitted detail lookup for that page.

TOOL CALL • STEP 10 • Field-read the retrieved drug page

CALL

```
tool_A inspect <artifact_1> /coarse/provider_results/provider_B/pages/0/text
```

RETURNED

exit_code: 0 tool: tool_A contract_version: the domain toolkit-tools/v1 stdout: {"schema_version": "tool_A/field-v1", "artifact_path": "<artifact_1>", "json_pointer": "/coarse/provider_results/provider_B/pages/0/text", "value_kind": "str", "value": "--- uid: <record_1> type: drug name: "dostarlimab" aliases: ["Dostarlimab", "Dostarlimab (Jemperli)"] n_papers: 238 n_claims: 388 n_claims_used: 13 grounded: false synth_model: "glm-5.2" prompt_template_version: "entity_page_v1" updated: 2026-07-06 tags: ["drug"]

dostarlimab

Overview

Dostarlimab is an FDA-approved anti-PD-1/PD-L1 antibody widely used in cancer treatment [ref-9], [ref-11]. It is an IgG4 isotype designed to avoid tumor-reactive T cell depletion, with little to no binding to Fc or complement protein C1q [ref-7]. Its pharmacokinetic profile is characterized by a 2-compartment model with time-dependent linear elimination, a mean terminal elimination half-life of 25.4 days, a mean clearance of 0.007 L/h, and a steady state volume of distribution of 5.3 L [ref-1], [ref-2]. As a tumor-agnostic therapy based on genetic changes, dostarlimab is highly relevant for drug repurposing across multiple histologies [ref-10].

Repurposing directions

Strongly supported directions

Dostarlimab is strongly supported as a treatment for [[dostarlimab → cancer]], with high-quality clinical trial and review evidence plus in vivo animal data indicating beneficial outcomes [ref-3]. It is strongly supported for [[dostarlimab → advanced solid tumors]], [[dostarlimab → solid tumors]], [[dostarlimab → dMMR solid tumors]], and [[dostarlimab → dMMR recurrent or advanced solid tumors]], with tissue-agnostic FDA approval for mismatch repair-deficient tumors [ref-10].

For endometrial malignancies, the drug is strongly supported across multiple specific presentations: [[dostarlimab → endometrial cancer]], [[dostarlimab → advanced endometrial cancer]], [[dostarlimab → advanced or recurrent EC]], [[dostarlimab → advanced EC]], [[dostarlimab → dMMR endometrial cancer]], [[dostarlimab → dMMR recurrent or advanced endometrial

cancer]], [[dostarlimab → recurrent or advanced endometrial cancer]], and [[dostarlimab → primary advanced or recurrent endometrial cancer]].

Dostarlimab shows exceptional promise in colorectal and rectal malignancies, strongly supported for [[dostarlimab → rectal cancer]], [[dostarlimab → locally advanced rectal cancer]], [[dostarlimab → dMMR rectal cancer]], and [[dostarlimab → colorectal cancer]], with clinical trials reporting 100% clinical complete responses in deficient mismatch repair patients.

It is also supported by moderate-to-high strength clinical-trial evidence for [[dostarlimab → NSCLC]], demonstrating durable responses across PD-L1 subgroups.

Directionally positive directions

[[dostarlimab → prostate cancer]] is directionally positive but constrained in scope. While dostarlimab is clinically approved for dMMR or high TMB solid tumors, qualifying biomarkers are rare in prostate cancer, though T cell signature enrichment in high-grade tumors supports biological plausibility.

Mechanisms & targets

Dostarlimab acts primarily by binding with high affinity (KD 300 pM) to human and cynomolgus monkey [[pd-1]], also known as [[programmed cell death protein 1]] [ref-5]. This binding causes conformational rearrangements in the BC, C'D, and FG loops of PD-1 [ref-6]. As a functional antagonist, it suppresses the interaction between PD-1 and [[pd-l1]] to support T cell activation [ref-4], [ref-9]. It also directly inhibits the function of [[pd-l2]] at the junction of PD-L2 and PD-1 to prevent immune escape of tumors [ref-13].

Mechanistically, this blockade restores T cell activity and enhances immune infiltration to suppress tumor growth, leading to increased IL-2 production in human CD4+ mixed lymphocyte reaction assays [ref-12]. The drug specifically targets [[mismatch repair-deficient]] (dMMR) or methylated tumors, where loss of mismatch-repair genes (such as MLH1, MSH6) predicts therapeutic sensitivity [ref-10].

Contradictions & open questions

While dostarlimab is approved for multiple indications, as of March 2023 there were no approved MHRA indications for seven specified cancer types [ref-4]. In [[dostarlimab → dMMR endometrial cancer]], PTEN mutations may identify a subset of patients with poor response, representing a caveat to the overall efficacy [ref-10]. Additionally, there is a notable gap in the approval cohort for [[dostarlimab → dMMR solid tumors]] regarding lung cancer patients, leaving efficacy in that specific histology uncharacterized.

Pharmacokinetically, there are no known drug-drug interactions as dostarlimab is not a cytochrome P450 or drug transporter substrate [ref-8].

See also

- [[pd-1]]
- [[pd-l1]]
- [[pd-l2]]
- [[programmed cell death protein 1]]
- [[FC]]
- [[mismatch repair-deficient]]

References

- [ref-9] Bio Protoc. 2024 Sep 5; 14(17):e5057 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/39282229)
- [ref-11] Sci Rep. 2026 Mar 28; 16:10773 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/41904276)
- [ref-7] Biomolecules. 2022 Jul 26; 12(8):1031 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/35892341)
- [ref-1] Biomolecules. 2022 Jul 26; 12(8):1031 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/35892341)
- [ref-2] Biosensors (Basel). 2022 Aug 8; 12(8):617 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/36005013)
- [ref-10] Front Pharmacol.; 17:1804164 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/0)
- [ref-3] J Immunol Res. 2026 Feb 16; 2026:7864229 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/41811823)
- [ref-5] Hum Vaccin Immunother.; 19(1):2178220 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/36762991)
- [ref-6] Biomolecules. 2022 Jul 26; 12(8):1031 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/35892341)
- [ref-4] BMJ Open. 2024 Feb 15; 14(2):e075981 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/38365286)
- [ref-13] Front Immunol. 2024 Mar 28; 15:1359532 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/38605944)
- [ref-12] Biosensors (Basel). 2022 Aug 8; 12(8):617 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/36005013)

- [ref-8] Biomolecules. 2022 Jul 26; 12(8):1031 --- [PubMed](https://pubmed.ncbi.nlm.nih.gov/35892341) ", "truncated_for_output": false, "max_chars": 12000}

{ "stderr": "" }

REASONING • STEP 11 • Decide to escalate the search depth

REASONING SUMMARY

Planning decision-relevant follow-up query Planning next-tier rerun with breast cancer

MODEL NARRATION

The first pass did not resolve the Stage II breast cancer anchor or build any paths. Because that is decision-critical coverage rather than a negative result, I'm using the allowed next-tier rerun into a new artifact to see whether the disease anchor and any grounded connections resolve.

TOOL CALL — FAILED • STEP 12 • Escalated search REJECTED — invalid tier argument

CALL

```
tool_A connect "Dostarlimab" "Stage II Breast Cancer" --tier deep --timeout 240 --output <artifact_2> --summary
```

RETURNED

```
exit_code: 2 tool: tool_A contract_version: the domain toolkit-tools/v1
stderr:
usage: tool_A connect [-h] [--methods METHODS] [--source {wiki,kg}]
                      [--max-page-chars MAX_PAGE_CHARS]
                      [--max-hints-per-type MAX_HINTS_PER_TYPE]
                      [--a-id A_ID] [--b-id B_ID]
                      [--a-type {drug,disease,gene,protein,pathway}]
                      [--b-type {drug,disease,gene,protein,pathway}]
                      [--tier {faster,standard,extra,max}]
                      [--seed {auto,llm,heuristic}] [--answer]
                      [--evidence] [--top-k TOP_K]
                      [--timeout TIMEOUT] [--no-merge] [--legacy]
                      [--stage2 {suggest,auto}] [--with-specialists]
                      [--max-candidate-paths MAX_CANDIDATE_PATHS]
                      [--max-coarse-depth MAX_COARSE_DEPTH]
                      [--max-deep-searches MAX_DEEP_SEARCHES]
                      [--specialist-top-k SPECIALIST_TOP_K]
                      [--specialist-max-hops SPECIALIST_MAX_HOPS]
                      [--specialist-max-expansions SPECIALIST_MAX_EXPANSIONS]
                      [--specialist-neighbor-limit SPECIALIST_NEIGHBOR_LIMIT]
                      [--specialist-timeout SPECIALIST_TIMEOUT]
                      [--pretty] [--output OUTPUT] [--summary]
                      [--stdin]
                      [a] [b]
tool_A connect: error: argument --tier: invalid choice: 'deep' (choose from faster, standard,
extra, max)
{
  "stdout": ""
}
```

REASONING • STEP 13 • Localise the failure to one bad argument of its own making

REASONING SUMMARY

Correcting invalid syntax to use max

MODEL NARRATION

The tier name I tried was invalid; the tool only exposes 'faster', 'standard', 'extra', and 'max'. I'm rerunning at the actual next tier, 'max', with the same bounded timeout and a new output path.

TOOL CALL • STEP 14 • Repaired reissue: same query, only the tier changed

CALL

```
tool_A connect "Dostarlimab" "Stage II Breast Cancer" --tier max --timeout 240 --output <artifact_3> --
summary
```

RETURNED

exit_code: 0 tool: tool_A contract_version: the domain toolkit-tools/v1 stdout: {"output_path": "<artifact_3>", "summary": {"schema_version": "tool_A/v2", "stage": "stage1", "workflow": {"review_before_next_call": true, "stage1_follow_up": "Optionally query one decision-relevant related_page_hint for more Stage-1 literature context.", "stage1_follow_up_command_template": "tool_A connect <hint.title> --a-id <hint.uid> --stage2 suggest --output <new-stage1-path> --summary", "stage2": "Only after reviewing Stage 1, optionally run one advertised ready expansion route when a deeper connection could change the conclusion."}, "artifact_inspection": {"default": "not_needed", "reason": "This projection already contains provider status, selected paths, load-bearing evidence, anchor and relationship pages, related-page hints, conflicts, and optional expansion candidates.", "exception": "Run only an item's supplied detail_command, once, when its truncated_for_summary flag is true and the omitted text is load-bearing. An absent relationship page is not a missing field and never authorizes artifact or filesystem search.", "provider_results_shape": "coarse.provider_results is an object keyed by canonical provider ID, not a list."}, "query": {"id": "query-<record>", "text": null, "entities": ["Dostarlimab", "Stage II Breast Cancer"], "input_entities": [{"id": "Dostarlimab", "name": "Dostarlimab", "type": "unknown", "species": null, "aliases": [], "resolution": {"source": "caller", "structured_input": true}}, {"id": "Stage II Breast Cancer", "name": "Stage II Breast Cancer", "type": "unknown", "species": null, "aliases": [], "resolution": {"source": "caller", "structured_input": true}}], "operation": "connect", "anchor_quality": {"trusted": false, "unresolved": [{"id": "<local_id>", "name": "stage ii breast cancer", "type": "unknown", "species": null, "aliases": ["Stage II Breast Cancer"], "resolution": {"source": "surface_form", "ambiguous": true, "providers": ["provider_A"], "provider_ids": {"provider_A": "stage ii breast cancer"}, "is_anchor": true, "input_aliases": ["Stage II Breast Cancer"]}}]}}, "resolved_entities": [{"id": "<record_1>", "name": "DOSTARLIMAB", "type": "drug", "species": null, "aliases": ["dostarlimab", "Dostarlimab"], "resolution": {"source": "provider_canonical_id", "ambiguous": false, "providers": ["provider_A", "provider_B"], "provider_ids": {"provider_A": "DOSTARLIMAB", "provider_B": "<record_1>"}, "is_anchor": true, "previous_ids": ["<local_id>"], "input_aliases": ["Dostarlimab"]}}, {"id": "<local_id>", "name": "stage ii breast cancer", "type": "unknown", "species": null, "aliases": ["Stage II Breast Cancer"], "resolution": {"source": "surface_form", "ambiguous": true, "providers": ["provider_A"], "provider_ids": {"provider_A": "stage ii breast cancer"}, "is_anchor": true, "input_aliases": ["Stage II Breast Cancer"]}}], "providers_run": [{"provider_id": "provider_A", "status": "partial", "algorithm_status": "partial", "algorithm_id": "weighted_best_first", "elapsed_ms": 108185.315, "timed_out": false, "budget_exhausted": true, "expansions": 400, "frontier_remaining": 8505, "early_stopped_on_connection": false, "effective_limits": {"max_expansions": 400, "neighbor_limit": 8, "max_depth": 6, "max_hops": 8, "top_k": 8, "timeout_seconds": 240}, "warnings": ["algorithm expansion budget exhausted after 400 frontier expansions; unvisited frontier remains"], "error": null}, {"provider_id": "provider_B", "status": "partial", "algorithm_status": "partial", "algorithm_id": "weighted_best_first", "elapsed_ms": 13514.596, "timed_out": false, "budget_exhausted": true, "expansions": 400, "frontier_remaining": 3173, "early_stopped_on_connection": false, "effective_limits": {"max_expansions": 400, "neighbor_limit": 8, "max_depth": 6, "max_hops": 8, "top_k": 8, "timeout_seconds": 240}, "warnings": ["algorithm expansion budget exhausted after 400 frontier expansions; unvisited frontier remains"], "error": null}], "projection_counts": {"candidate_paths": {"shown": 0, "total": 0}, "connections": {"shown": 0, "total": 0}, "load_bearing_evidence": {"shown": 0, "total_selected_edge_ids": 0}, "relationship_pages": {"shown": 0, "total": 0, "direct_for_anchor_pair_available": false}, "anchor_pages": {"shown": 1, "resolved_anchors": 1}, "related_page_hints": {"shown": 7, "total": 21}, "expansion_candidates": {"shown": 0, "total": 0}}, "candidate_paths": [], "expansion_candidates": [], "connections": [], "load_bearing_evidence": [], "anchor_pages": [{"uid": "<record_1>", "title": "dostarlimab", "type": "drug", "path": "drugs/dostarlimab.md", "n_papers": 238, "pmids": ["39282229", "41904276", "35892341", "36005013", "0", "41811823", "36762991", "38365286", "38605944"], "text": "--- uid: <record_1> type: drug name: "dostarlimab" aliases: ["Dostarlimab", "Dostarlimab (Jemperli)"] n_papers: 238 n_claims: 388 n_claims_used: 13 grounded: false synth_model: "glm-5.2" prompt_template_version: "entity_page_v1" updated: 2026-07-06 tags: ["drug"]

dostarlimab

Overview

Dostarlimab is an FDA-approved anti-PD-1/PD-L1 antibody widely used in cancer treatment [ref-9], [ref-11]. It is an IgG4 isotype designed to avoid tumor-reactive T cell depletion, with little to no binding to Fc or complement protein C1q [ref-7]. Its pharmacokinetic profile is characterized by a 2-compartment model with time-dependent linear elimination, a mean terminal elimination half-life of 25.4 days, a mean clearance of 0.007 L/h, and a steady state volume of distribution of 5.3 L [ref-1], [ref-2]. As a tumor-agnostic therapy based on genetic changes, dostarlimab is highly relevant for drug repurposing across multiple histologies [ref-10].

Repurposing directions

Strongly supported directions

Dostarlimab is strongly supported as a treatment for [[dostarlimab → cancer]], with high-quality clinical trial and review evidence plus in vivo animal data indicating beneficial", "truncated_for_summary": true, "full_text_artifact_json_pointer"

: "/coarse/provider_results/provider_B/pages/0/text", "detail_command": "tool_A inspect <artifact_3> /coarse/provider_results/provider_B/pages/0/text"}], "relationship_pages": [], "related_page_hints": [{"uid": "<record>", "title": "dostarlimab → endometrial cancer", "path": "hypotheses/dostarlimab → endometrial cancer.md", "n_papers": 29, "type": "hypothesis", "follow_up_stage": "stage1"}, {"uid": "<record_2>", "title": "PD-1", "path": "targets/PD-1.md", "n_papers": 12098, "type": "target", "follow_up_stage": "stage1"}, {"uid": "<record_1>", "title": "FC", "path": "drugs/FC.md", "n_papers": 20, "type": "drug", "follow_up_stage": "stage1"}, {"uid": "<record>", "title": "dostarlimab → locally advanced rectal cancer", "path": "hypotheses/dostarlimab → locally advanced rectal cancer.md", "n_papers": 12, "type": "hypothesis", "follow_up_stage": "stage1"}, {"uid": "<record_2>", "title": "PD-L1", "path": "targets/PD-L1.md", "n_papers": 10947, "type": "target", "follow_up_stage": "stage1"}, {"uid": "<record>", "title": "dostarlimab → rectal cancer", "path": "hypotheses/dostarlimab → rectal cancer.md", "n_papers": 11, "type": "hypothesis", "follow_up_stage": "stage1"}, {"uid": "<record_2>", "title": "programmed cell death protein 1 (PD-1)", "path": "targets/programmed cell death protein 1 (PD-1).md", "n_papers": 195, "type": "target", "follow_up_stage": "stage1"}], "stage2": {"policy": "suggest", "optional": true, "suggested_expansions": 0, "requested_routes": 0, "guidance": "Stage 1 is complete. The calling agent may answer now or request one or more compatible expansion routes only when a load-bearing weak, missing, or conflicting link could change the conclusion."}, "report": {"summary": "No complete connection path was reconstructed for Dostarlimab ↔ Stage II Breast Cancer. Provider results and unresolved gaps are retained for diagnosis.", "support_grade": "incomplete", "conflicts": [], "conflict_count": 0, "unresolved_gaps": [{"gap_id": null, "candidate_path_id": null, "subject": "<local_id>", "object": null, "required_provider": null, "reason": "unresolved_or_local_query_anchor"}], "unresolved_gap_count": 1, "optional_expansions": {"policy": "suggest", "suggested": 0, "requested": 0}, "diagnostics": {"non_load_bearing_conflict_count": 1398}, "interpretation": "Provider-native scores are shown separately and were not averaged. A support grade describes evidence structure, not treatment efficacy."}, "graph_warnings": []}}

{ "stderr": "" }

REASONING • STEP 15 • Direct pair exhausted; pivot to the mechanism target

REASONING SUMMARY

Planning related page queries Deciding PD-L1 page query

MODEL NARRATION

The max-tier rerun still leaves the disease unresolved and shows no candidate paths. One related target page could affect confidence in biological plausibility, so I'm querying the PD-L1 hint once for any literature context connecting the checkpoint axis to breast cancer.

TOOL CALL • STEP 16 • Follow the tool-suggested related-entity hint

CALL

```
tool_A connect "PD-L1" --a-id <record_2> --stage2 suggest --output <artifact_4> --summary
```

RETURNED

exit_code: 0 tool: tool_A contract_version: the domain toolkit-tools/v1 stdout: {"output_path": "<artifact_4>", "summary": {"schema_version": "tool_A/v2", "stage": "stage1", "workflow": {"review_before_next_call": true, "stage1_follow_up": "Optionally query one decision-relevant related_page_hint for more Stage-1 literature context.", "stage1_follow_up_command_template": "tool_A connect <hint.title> --a-id <hint.uid> --stage2 suggest --output <new-stage1-path> --summary", "stage2": "Only after reviewing Stage 1, optionally run one advertised ready expansion route when a deeper connection could change the conclusion."}, "artifact_inspection": {"default": "not_needed", "reason": "This projection already contains provider status, selected paths, load-bearing evidence, anchor and relationship pages, related-page hints, conflicts, and optional expansion candidates.", "exception": "Run only an item's supplied detail_command, once, when its truncated_for_summary flag is true and the omitted text is load-bearing. An absent relationship page is not a missing field and never authorizes artifact or filesystem search.", "provider_results_shape": "coarse.provider_results is an object keyed by canonical provider ID, not a list."}, "query": {"id": "query-<record>", "text": null, "entities": ["PD-L1"], "input_entities": [{"id": "<record_2>", "name": "PD-L1", "type": "unknown", "species": null, "aliases": [], "resolution": {"source": "caller", "structured_input": true}}], "operation": "discover", "anchor_quality": {"trusted": true, "unresolved": []}}, "resolved_entities": [{"id": "<record_2>", "name": "PD-L1", "type": "gene", "species": "Homo sapiens", "aliases": [], "resolution": {"source": "provider_canonical_id", "ambiguous": false, "providers": ["provider_B"], "provider_ids": {"provider_B": "<record_2>"}, "is_anchor": true, "input_aliases": ["PD-L1"]}}], "providers_run": [{"provider_id": "provider_A", "status": "unresolved", "algorithm_status": "unresolved", "algorithm_id": "weighted_best_first", "elapsed_ms": 11354.726, "timed_out": false, "budget_exhausted": false, "expansions": 0, "frontier_remaining": 0, "early_stopped_on_connection": false, "effective_limits": {"max_expansions": 150, "neighbor_limit": 5, "max_depth": 5, "max_hops": 7, "top_k": 5, "timeout_seconds": 240}, "warnings": [], "error": null}, {"provider_id": "provider_B", "status": "partial", "algorithm_status": "partial", "algorithm_id": "weighted_best_first", "elapsed_ms": 11074.581, "timed_out": false, "budget_exhausted": true, "expansions": 150, "frontier_remaining": 638, "early_stopped_on_connection": false, "effective_limits": {"max_expansions"

: 150, "neighbor_limit": 5, "max_depth": 5, "max_hops": 7, "top_k": 5, "timeout_seconds": 240}, "warnings": ["algorithm expansion budget exhausted after 150 frontier expansions; unvisited frontier remains"], "error": null}], "projection_counts" : {"candidate_paths": {"shown": 0, "total": 0}, "connections": {"shown": 0, "total": 0}, "load_bearing_evidence": {"shown": 0, "total_selected_edge_ids": 0}, "relationship_pages": {"shown": 0, "total": 0, "direct_for_anchor_pair_available": false}, " anchor_pages": {"shown": 1, "resolved_anchors": 1}, "related_page_hints": {"shown": 8, "total": 25}, "expansion_candidates": { "shown": 0, "total": 0}}, "candidate_paths": [], "expansion_candidates": [], "connections": [], "load_bearing_evidence": [], "anchor_pages": [{"uid": "<record_2>", "title": "PD-L1", "type": "target", "path": "targets/PD-L1.md", "n_papers": 10947, " pmids": ["36969168", "38565806", "27764774", "41693025", "32509787", "41059490", "36564129", "34759534", "40251364", "38254708", "39962074", "36231138", "33014787"], "text": "--- uid: <record_2> type: target name: "PD-L1" aliases: ["Pd-l1", " PD-L1 (combined positivity score)", "PD-L1 (CD274)", "Pd-L1", "pd-l1", "PD-L1 (CPS=105)", "PD-L1 (programmed death-ligand 1)", "Pd-l1 (Cd274)", "PD-L1 (B7-H1)"] n_papers: 10947 n_claims: 20406 n_claims_used: 400 grounded: false synth_model: " glm-5.2" prompt_template_version: "entity_page_v1" updated: 2026-07-06 tags: ["target"]

PD-L1

Overview

PD-L1 (programmed death-ligand 1), also known as CD274 or B7-H1, is an immune checkpoint protein and a critical target in oncology and drug repurposing. PD-L1 expression (measured via combined positivity score, tumor proportion score, or tumor cell/immune cell algorithms) frequently predicts survival and response to immune checkpoint inhibitors across multiple advanced malignancies [ref-248-257, ref-277, ref-294, ref-297]. While its primary clinical application is in cancer therapy, non-traditional repurposing approaches---such as managing sepsis with BMS-936559 [ref-298] or using 89Zr-labelled atezolizumab as an imaging tracer to predict clinical response [ref-220]---highlight its expanding therapeutic footprint.

Repurposing directions

PD-L1 blockade de", "truncated_for_summary": true, "full_text_artifact_json_pointer": "/coarse/provider_results/provider_B/pages/0/text", "detail_command": "tool_A inspect <artifact_4> /coarse/provider_results/provider_B/pages/0/text"}], " relationship_pages": [], "related_page_hints": [{"uid": "<record>", "title": "atezolizumab → triple-negative breast cancer" , "path": "hypotheses/atezolizumab → triple-negative breast cancer.md", "n_papers": 221, "type": "hypothesis", "follow_up_ stage": "stage1"}, {"uid": "<record_2>", "title": "IFNγ", "path": "targets/IFNγ.md", "n_papers": 777, "type": "target", "follow_ up_stage": "stage1"}, {"uid": "disease:<record>", "title": "cancer", "path": "diseases/cancer.md", "n_papers": 71479, "type": " disease", "follow_up_stage": "stage1"}, {"uid": "<record_1>", "title": "pembrolizumab", "path": "drugs/pembrolizumab.md", "n_ papers": 8116, "type": "drug", "follow_up_stage": "stage1"}, {"uid": "mechanism:<record>", "title": "PD-L1 blockade", "path": " mechanisms/PD-L1 blockade.md", "n_papers": 88, "type": "mechanism", "follow_up_stage": "stage1"}, {"uid": "<record>", " title": "atezolizumab → hepatocellular carcinoma", "path": "hypotheses/atezolizumab → hepatocellular carcinoma.md", "n_ papers": 145, "type": "hypothesis", "follow_up_stage": "stage1"}, {"uid": "<record_1>", "title": "nivolumab", "path": "drugs/ nivolumab.md", "n_papers": 6147, "type": "drug", "follow_up_stage": "stage1"}, {"uid": "<record>", "title": "nivolumab → renal cell carcinoma", "path": "hypotheses/nivolumab → renal cell carcinoma.md", "n_papers": 71, "type": "hypothesis", "follow_up_ stage": "stage1"}], "stage2": {"policy": "suggest", "optional": true, "suggested_expansions": 0, "requested_routes": 0, "guidance": "Stage 1 is complete. The calling agent may answer now or request one or more compatible expansion routes only when a load-bearing weak, missing, or conflicting link could change the conclusion."}, "report": {"summary": "No complete connection path was reconstructed for PD-L1. Provider results and unresolved gaps are retained for diagnosis.", "support_grade": "incomplete", " conflicts": [], "conflict_count": 0, "unresolved_gaps": [], "unresolved_gap_count": 0, "optional_expansions": {"policy": "suggest", "suggested": 0, "requested": 0}, "diagnostics": {"non_load_bearing_conflict_count": 226}, "interpretation": "Provider-native scores are shown separately and were not averaged. A support grade describes evidence structure, not treatment efficacy."}, " graph_warnings": []}}

{ "stderr": "" }

REASONING • STEP 17 • Classify the hint as class-adjacent, not direct evidence

REASONING SUMMARY

Evaluating detail_command usage limits Scoring dostarlimab relevance for breast cancer

MODEL NARRATION

I have enough bounded evidence to make a calibrated call: strong checkpoint-mechanism plausibility, but no resolved stage-II breast-cancer anchor or direct dostarlimab-breast-cancer path in the sanctioned KG summaries. Submitting now.

SUBMISSION • STEP 18 • Submit

FINAL ANSWER

promisingness 0.3 **confidence** 0.48

Dostarlimab is a validated anti-PD-1/PD-L1 checkpoint antibody with tissue-agnostic and multiple-cancer activity, including dMMR solid tumors, so the mechanism is biologically plausible in immunogenic breast-cancer subsets. However, the mandated tool_A searches did not resolve Stage II Breast Cancer as a grounded disease anchor and reconstructed no candidate paths, connections, relationship pages, or load-bearing evidence for Dostarlimab to Stage II Breast Cancer even at the max tier. A PD-L1 related-page query supports checkpoint relevance in oncology and surfaced triple-negative breast cancer evidence for another PD-L1 inhibitor, but that is class-adjacent rather than dostarlimab-specific or stage-II-specific. I therefore score the pair below Phase-2-success territory but above purely preclinical, reflecting plausible repurposing with no sanctioned direct clinical support for this exact pair.

KEY CLAIMS (7)

- Dostarlimab was resolved as a drug anchor, but Stage II Breast Cancer remained unresolved or ambiguous in the Stage-1 search.
  `source: tool:tool_A connect Dostarlimab + Stage II Breast Cancer --tier extra`
  support: resolved_entities included DOSTARLIMAB as type drug, while anchor_quality listed unresolved <local_id> with reason/source surface_form and ambiguous true.
- No complete Dostarlimab to Stage II Breast Cancer connection was reconstructed in the bounded Stage-1 search.
  `source: tool:tool_A connect Dostarlimab + Stage II Breast Cancer --tier extra`
  support: report.summary: No complete connection path was reconstructed for Dostarlimab ↔ Stage II Breast Cancer; candidate_paths, connections, load_bearing_evidence, and relationship_pages were all empty.
- The max-tier rerun still found no direct path or relationship evidence for the exact pair, leaving the disease anchor unresolved.
  `source: tool:tool_A connect Dostarlimab + Stage II Breast Cancer --tier max`
  support: report.summary: No complete connection path was reconstructed for Dostarlimab ↔ Stage II Breast Cancer; unresolved_gaps reason: unresolved_or_local_query_anchor; candidate_paths shown 0 total 0 and connections shown 0 total 0.
- Dostarlimab is an FDA-approved anti-PD-1/PD-L1 antibody with cancer-treatment use and tumor-agnostic relevance.
  `source: PMID:39282229; PMID:41904276; PMID:0 via <record_1> page`
  support: Overview: Dostarlimab is an FDA-approved anti-PD-1/PD-L1 antibody widely used in cancer treatment [ref-9], [ref-11]... As a tumor-agnostic therapy based on genetic changes, dostarlimab is highly relevant for drug repurposing across multiple histologies [ref-10].
- Dostarlimab has strong support in dMMR solid tumors and several non-breast cancers, including endometrial and rectal/colorectal settings, but the retrieved drug page did not list breast cancer as a supported direction.
  `source: <record_1> page inspected by tool_A inspect <artifact_1> /coarse/provider_results/provider_B/pages/0/text`
  support: Strongly supported directions listed cancer, advanced solid tumors, dMMR solid tumors, endometrial malignancies, rectal cancer, locally advanced rectal cancer, colorectal cancer, and NSCLC; no breast cancer direction appeared in that section.
- The mechanism is checkpoint blockade: dostarlimab binds PD-1 and suppresses PD-1 interactions with PD-L1/PD-L2 to restore T-cell activity.
  `source: PMID:36762991; PMID:38365286; PMID:38605944; PMID:36005013 via <record_1> page`
  support: Dostarlimab acts primarily by binding with high affinity (KD 300 pM) to human and cynomolgus monkey pd-1... As a functional antagonist, it suppresses the interaction between PD-1 and pd-l1 to support T cell activation... inhibits... PD-L2 and PD-1... blockade restores T cell activity and enhances immune infiltration.
- PD-L1 is a broad oncology checkpoint target and the related-page hints surfaced class-adjacent triple-negative breast cancer evidence for atezolizumab, supporting plausibility but not direct dostarlimab evidence.
  `source: tool:tool_A connect PD-L1 --a-id <record_2>`
  support: PD-L1 Overview: PD-L1... is an immune checkpoint protein and a critical target in oncology... related_page_hints included title: atezolizumab → triple-negative breast cancer, n_papers: 221.

D.3. Process score

Table 38: TRACES process score with all sub-rubrics. A capability is the weighted mean of its sub-rubric bands (0/1/2) rescaled to $[0, 4]$. A sub-rubric marked CRITICAL caps its capability when it falls below the top band. NR marks a sub-rubric the trajectory gave no opportunity to exercise; it is excluded from the mean rather than scored zero.

C	**Long-horizon State Coherence**	**3.00** / 4

Table 38 (continued)

The assigned pair is held from task intake through both searches to submission; the final answer is built from what was retrieved rather than reasoning around it. Two bands are lost: the supplied identifiers were available but were not used to anchor the search, and the reported confidence is not tied to a named retrieved quantity.

2	Cross-step integration	The final judgement traceably incorporates the main tool_A returns: the max-tier query's lack of a reconstructed connection and zero candidate paths are explicitly used to lower the promisingness, while the PD-L1 hint is used only as class-adjacent plausibility rather than direct evidence. The submission also lists key claims tying each conclusion to the extra/max queries, the inspected drug page, and the PD-L1 related-page query, so retrieved evidence entries are folded into the final score rather than ignored.
2	State & constraint retention CRITICAL	The trajectory consistently centers the supplied pair from task inspection through the main tool_A calls, using sanctioned offline the domain toolkit tooling for the exact Dostarlimab + Stage II Breast Cancer query. The final status claim that no direct connection was found is directly retrieved from the max-tier sanctioned result. Although the agent queries PD-L1 and notes atezolizumab/triple-negative breast cancer as plausibility context, it explicitly does not present that as direct evidence for the asked pair, calling it class-adjacent rather than dostarlimab- or stage-II-specific.
1	Task decomposition	The trajectory had a real dependency-ordered workflow: it inspected the task, made sanctioned tool_A calls, inspected/expanded evidence, and only then submitted. But it does not reach the score-2 anchor requirement because the supplied <drug_id_ns>/<disease_id_ns> IDs were available yet the connect call used surface names, and the disease anchor remained LOCAL/unresolved; the final relies on recognizing that unresolved state rather than first anchoring the pair correctly by drug_id/disease_id. This fits score 1 rather than 0 or 2.
1	Terminal coherence	I agree with the reviewer that t=18 contains a strong qualitative contrast: biologically plausible mechanism/class evidence versus no sanctioned direct support for the exact pair, and that this is a spelled-out evidence-category-to-band mapping for the 0.3 promisingness. I also agree the key_claims trace the rationale to concrete retrieved entries, including no-path/unresolved-anchor findings and class-adjacent PD-L1 breast-cancer evidence. However, the score-2 anchor specifically requires promisingness, confidence, and rationale each trace back with the evidence-strength-to-value mapping spelled out. A qualitative evidence-category-to-band justification can be sufficient; exact numeric calibration to 0.48 is not inherently required. But here the final never explicitly connects those categories to confidence as opposed to promisingness/rationale. The confidence value is present, but the text does not say that confidence is intermediate because evidence is tool-grounded yet indirect/incomplete. Thus the concession that the mapping to 0.48 was not spelled out remains material and is priced into a score of 1.

E Evidence Fidelity **3.50** / 4

Strongest dimension. Every key claim names a resolvable retrieved record and its support is traceable to that same record; retrieved fact, own inference, and acknowledged unknown are kept apart in the prose — the mechanism is called plausible*, not established, and the class-level hint is explicitly labelled class-adjacent. One band is lost because load-bearing evidence is not accompanied by an explicit recency statement.*

2	Claim grounding	The submitted key_claims consistently identify resolvable tool commands or inspected pages as sources, and their supports are traceable to those same records. For example, the unresolved anchor claim in the final is supported by the same Stage-1 output showing the local unresolved breast-cancer anchor, and the no-connection claim matches the Stage-1 report text. The mechanistic and tumor-agnostic claims are directly supported by the inspected dostarlimab page text. Even the PD-L1/breast-cancer support is careful: it cites the PD-L1 query and characterizes the hint as class-adjacent rather than direct evidence, avoiding overstating the record.
2	Fact / inference / unknown separation	The final answer cleanly distinguishes source-backed findings from inference and unknowns. It labels the mechanistic extension as biologically plausible, not established; explicitly states the tool_A limitation that the disease anchor was not resolved; calls PD-L1/TNBC evidence class-adjacent rather than direct; and frames the score as plausible repurposing with no sanctioned direct clinical support for the exact pair. I did not find a load-bearing conjecture presented as a settled tool_A fact.

Table 38 (continued)

2	Source integrity CRITICAL	The submitted answer's load-bearing negative claims point to returned named fields from the tool_A output, including zero candidate paths/connections/load-bearing evidence in step 14. Positive mechanism/context claims cite locatable returned records: PMIDs and the drug page uid/artifact path from returned/inspected pages. The PD-L1 class-adjacent claim cites the actual PD-L1 connect call and returned hints. I found no load-bearing item cited merely as vague 'literature' or 'KG' support.
1	Temporal-leakage discipline	The trajectory stayed within the sanctioned offline environment: bash was used to read local toolkit/access files and the main evidence gathering used tool_A, with the access file noting the the sandbox has no network. However, load-bearing claims such as the final FDA/checkpoint-antibody rationale were stated as current without any explicit recency handling. The retrieved drug page carried an updated stamp (2026-07-06), but the submission did not state that status or caveat possible staleness for updated/undated synthesized pages. Thus there is no score-0 leakage breach, but the required recency discipline for score 2 is missing.

A Hypothesis Management **3.00** / 4

A concrete testable mechanism is stated with named targets and a cellular effect, and the negative retrieval is retained in the final answer and used to lower the score rather than being discarded. Two bands are lost: only one biological mechanism is advanced where the sub-rubric asks for a competing no-benefit hypothesis to be set beside it, and no explicit before/after revision of the estimate is shown as the negative evidence arrived.

1	Alternative paths considered	On revision, the reviewer's point is persuasive: the agent mainly advances one biological mechanism, PD-1/PD-L1 checkpoint blockade, and the statement that the tool found no candidate paths or load-bearing evidence is an evidence-gap statement, not an explicitly formulated competing biological 'no benefit' or 'symptomatic-only' hypothesis. The final does qualify the single mechanism by noting class-adjacent and non-specific evidence, so it acknowledges weakness/other possibilities, but it never places two mechanistic paths side by side nor explicitly argues that the drug would have no benefit beyond symptomatic-only care. Therefore this fits score 1 rather than 2.
2	Contradiction retention	The trace offered a clear null/contradictory retrieval: max-tier tool_A found no complete path and labeled support incomplete. The submission retained this in the rationale, noting the unresolved disease anchor and no direct evidence, and explicitly used it to lower the deliverable promisingness/confidence by scoring below Phase-2-success territory while still acknowledging mechanistic plausibility/class-adjacent support.
2	Hypotheses made explicit	The final deliverable explicitly states a testable checkpoint-blockade mechanism with named targets (PD-1/PD-L1, plus PD-L2 in key claims), a cellular process (restored T-cell activity), and relates it to the disease context through immunogenic breast-cancer subsets. Although it caveats that evidence for exact Stage II Breast Cancer is indirect, the mechanism itself is concrete and disease-relevant, meeting the high band.
1	Hypothesis update on evidence	The trajectory maps retrieved records into the final score: the max-tier search found no complete path (t14), while the final rationale explicitly uses that negative evidence plus the PD-L1 related-page evidence to justify a below-Phase-2 but above-preclinical score (t18). However, I found no visible prior, baseline, interim score, or explicit upward/downward update caused by a named retrieved record; the score is a static synthesis rather than a demonstrated ranking movement. This matches score 1 rather than 2.

S Boundary & Failure Reasoning **2.50** / 4

Weakest dimension, and the honest reading of it. Research-use framing is maintained throughout and the evidence gap for this exact pair is stated plainly. But the counter-case is argued almost entirely as an absence of evidence*: the submission does not name a concrete biological or clinical reason this repurposing could fail in this setting, does not raise an indication-specific safety or feasibility constraint, and never explains why the stated confidence value is the appropriate one.*

2	Boundary & scope CRITICAL	The submission stays within research-use framing for the exact drug-disease pair: it describes mechanistic plausibility and explicitly qualifies the evidence as lacking direct clinical support for Stage II Breast Cancer. Its key claims state that no complete connection was reconstructed and that PD-L1/TNBC evidence is only class-adjacent, matching the retrieved incomplete support structure. I found no dosing, administration, monitoring, or recommendation to start/switch/treat patients with dostarlimab for this population.

Table 38 (continued)

1	Failure modes	On revision, I agree the cited counter-evidence is primarily an evidence-gap argument, not a concrete biological/clinical failure mechanism. The final submission says the mechanism is plausible in immunogenic breast-cancer subsets, but it does not state that Stage II tumors may lack immunogenicity, that dostarlimab may have insufficient target engagement, inadequate PK/exposure, limited tissue delivery, or a wrong endpoint/population mechanism causing failure. Its concerns that the evidence is class-adjacent, not dostarlimab- or stage-II-specific, and that there is no direct clinical support amount to limited-evidence hedging. Therefore this fits score 1 rather than score 2.
1	Safety & feasibility gating	The trajectory had some generic feasibility/safety-adjacent material in the inspected drug page, including pharmacokinetics and no known drug-drug interactions, but it did not name an indication-specific safety or feasibility constraint for Stage II Breast Cancer. The submitted rationale and score were driven by lack of direct evidence and class-adjacent plausibility, not tolerability, contraindications, target-organ toxicity, administration, or monitoring burden.
1	Uncertainty calibration	The submission names uncertainty sources (unresolved disease anchor, no paths/load-bearing evidence, class-adjacent PD-L1 evidence) satisfying the score-1 prerequisite. However, the rationale never mentions 'confidence' and never explains why confidence_score=0.48 is appropriate. The only 'therefore score' sentence references promisingness tier anchors ('below Phase-2-success territory but above purely preclinical'), which maps to the 0.3 promisingness value, not the 0.48 confidence value. The task instructs 'Report low confidence when evidence is thin' yet 0.48 is moderate, not low, and this choice receives no justification. Score 2 requires explaining a confidence_score consistent with described evidence strength; merely naming uncertainty without connecting it to the numerical confidence value is score 1.

T Tool Use & Execution State Management **3.50** / 4

Manuals were read before any retrieval, the mandated bounded search was issued in the documented form, returned artefacts were opened and their content reused in the final answer, and search depth was escalated once and then stopped with an explicit sufficiency judgement. One band is lost on call correctness, for the avoidable invalid-parameter attempt at step 12 — the same event that makes Repair gradeable.

1	Call correctness	The agent did read the workflow/help first and its initial required Stage-1 call used the documented command form correctly. However, it later made an avoidable argument error by inventing/using an unsupported tier 'deep', which the tool rejected, then patched by retrying with the valid 'max' tier. Also the pair was passed by surface names rather than the task's supplied IDs, and the disease resolved as local/unknown, so this does not meet the high-band exact anchor/argument standard. The calls were ultimately usable, fitting score 1 rather than 0 or 2.
2	Cost & stopping discipline	The agent did not stop after a single weak return: it escalated from an extra-tier pair query to max tier, inspected the truncated drug page, and made one relevant PD-L1 hint query. It then submitted a low/moderate confidence conclusion with an explicit sufficiency judgment: the exact pair still had no reconstructed path or load-bearing evidence even at max tier, while the PD-L1 breast-cancer evidence was only class-adjacent, hence the stated below-Phase-2 score. There is a minor non-incremental invalid deep-tier attempt, but it immediately self-corrected to max and converged rather than looping.
2	Result absorption	The agent opened and read returned artifacts, including the full drug page and concrete summary fields, then made those details resurface in key_claims/source/support. Because the search returned no paths or edges, there were no edge directions/confidences to absorb; nevertheless it used specific returned items such as empty candidate_paths/connections/load_bearing_evidence, unresolved/no complete path reporting, drug-page mechanism/approval text, and PD-L1 related-page hint details to support the final judgment. This exceeds count/headline-only absorption.
2	Tool selection CRITICAL	The agent used retrieval rather than recall and chose tool_A commands aligned with the open questions. It began with the required pairwise connect for Dostarlimab and Stage II Breast Cancer, then the returned summary showed no suggested expansions, making skipped expand acceptable. It later followed a decision-relevant related-page hint for PD-L1 using the correct target id and stage2 suggest form. The invalid deep-tier attempt is a parameter/call-correctness problem, not evidence that the subcommand choice was unrelated to the pair.

R Self-correction under Verification **4.00** / 4

Table 38 (continued)

This is why the case was selected. *A real failure occurred at step 12 and was repaired at step 14, so this capability is genuinely* rated *rather than skipped. Four of its five sub-rubrics score at the top band: the rejection was read as a self-made bad argument rather than a tool or data fault; the fix changed* only *the offending parameter and resumed the same evidence line; the repaired call's substantive content was then actually used, not merely checked for success; and the same error class never recurred. The fifth is* NR*: with the error corrected on the very next action, the trace never exhibited a* sustained *localisation problem to grade separately.*

NR	Error localisation	The rubric's trigger condition requires 'a non-ok/rejected tool status, an explicit error, or an empty/ negative result the agent had relied on' and explicitly excludes 'A call that succeeded (exit 0)'. The only incident was t12's invalid --tier deep flag, which the agent immediately corrected to --tier max at t14. The connection calls at t8 and t14 both succeeded with exit_code 0 and returned structured partial results with budget_exhausted: true, error: null, timed_out: false. The task brief explicitly frames such partial results as 'usable bounded coverage---not a provider failure or negative evidence.' The unresolved disease anchor appeared in anchor_quality.unresolved with documented reason 'unresolved_or_local_query_anchor', but this was an operational limitation in a successful API response, not a repairable tool-call error. The documented workflow in <manual_ B> instructs entity names with no mention of <disease_id_ns> IDs or --b-id usage for disease resolution, and the KG provider uses internal hash IDs (disease:<record> format) not external ontology identifiers. There is no evidence that --b-id <disease_id> would have resolved the anchor. Because the connection calls succeeded and returned documented partial coverage per the API design, no qualifying repairable incident occurred under the rubric's definition.
2	Failure comprehension	The trajectory contains a clear repairable incident: the agent tried an unsupported tier, and the tool explained that 'deep' was an invalid choice with valid options including max. The next action selected the valid max tier, showing the agent read the failure as a bad argument rather than tool/ data failure. The empty/no-path results were also represented in the deliverable as unresolved coverage with zero candidate paths and connections, not overstated as definitive negative evidence.
2	No repeated error	There was a clear repairable incident: the agent attempted an unsupported tier, and the tool reported that 'deep' was invalid while listing allowed tiers. The next analogous call changed the argument to '--tier max' and succeeded with exit_code 0. I found no recurrence of the same invalid tier error afterward, so the error class never returned.
2	Re-verification	The trajectory had a qualifying incident: the attempted deep tier produced an explicit invalid-choice error. The agent substituted the valid max tier for the same Dostarlimab/Stage II Breast Cancer connection query, and the returned record was not merely checked for success; its substantive result (no reconstructed connection, unresolved disease anchor, zero paths/connections) was used in the final rationale/key claims. The failed deep-tier line was not carried forward as evidence.
2	Targeted repair	The trace contains a clear repairable incident: the agent attempted the same tool_A connect line with an unsupported --tier deep, and the tool reported the valid choices. The next relevant action made the minimal root-cause fix by changing only the invalid tier to the supported max tier while keeping the same Dostarlimab/Stage II Breast Cancer evidence line, and that repaired command succeeded.