STELLA: Self-Evolving LLM Agent for Biomedical Research

Ruofan Jin[1*], Zaixi Zhang[1*†], Mengdi Wang[1], Le Cong[2]

[1]Princeton University, Princeton, NJ, USA

[2]Stanford University, Stanford, CA, USA

[*]Co-first authors [†]Corresponding author: zz8680@princeton.edu

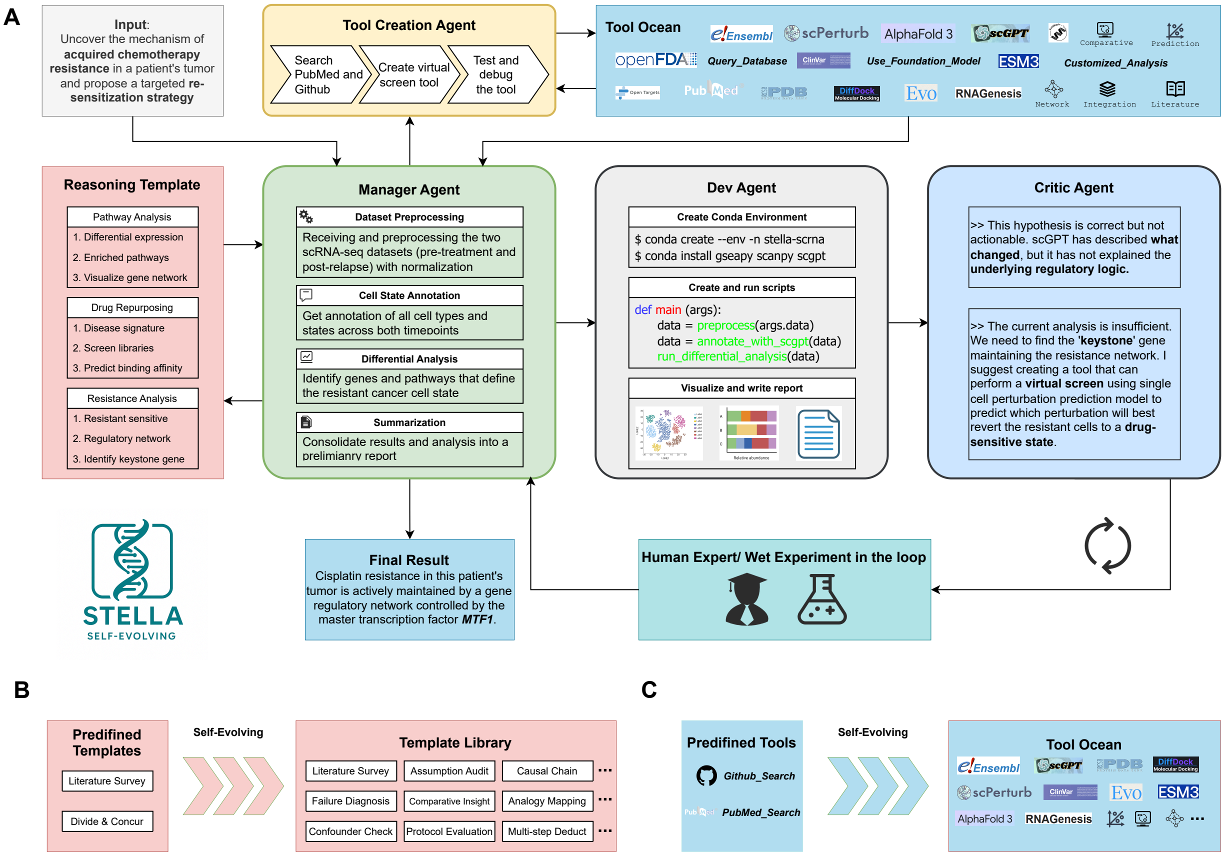
The rapid growth of biomedical data, tools, and literature has created a fragmented research landscape that outpaces human expertise. While AI agents offer a solution, they typically rely on static, manually-curated toolsets, limiting their ability to adapt and scale. Here, we introduce STELLA, a self-evolving AI agent designed to overcome these limitations. STELLA employs a multi-agent architecture that autonomously improves its own capabilities through two core mechanisms: an evolving Template Library for reasoning strategies and a dynamic Tool Ocean that expands as a Tool Creation Agent automatically discovers and integrates new bioinformatics tools. This allows STELLA to learn from experience. We demonstrate that STELLA achieves state-of-the-art accuracy on a suite of biomedical benchmarks, scoring approximately 26% on Humanity's Last Exam: Biomedicine, 54% on LAB-Bench: DBQA, and 63% on LAB-Bench: LitQA, outperforming leading models by up to 6 percentage points. More importantly, we show that its performance systematically improves with experience; for instance, its accuracy on the Humanity's Last Exam benchmark almost doubles with increased trials. STELLA represents a significant advance towards AI Agent systems that can learn and grow, dynamically scaling their expertise to accelerate the pace of biomedical discovery.

Introduction

Modern biomedical research is defined by both immense opportunity and staggering complexity. As a cornerstone of science, it generates vast quantities of data from large-scale experiments, but this progress is hampered by a research landscape that is profoundly fragmented (1–3). The knowledge, specialized software, and databases required to make discoveries are numerous, constantly evolving, and dispersed, forcing researchers to expend significant time and effort on the manual and labor-intensive task of discovering, learning, and integrating these disparate resources. While the advent of AI agents holds the promise of automating this intricate work (4–6), current systems inherit a critical limitation: they typically rely on manually curated, static toolsets (7–14). This approach is inefficient, fails to scale, and cannot keep pace with the rapid evolution of biomedical science, leaving the agents perpetually behind the cutting edge. This raises a critical question: Can we design a self-evolving agent that transcends these limitations by automatically discovering and integrating new tools, continuously updating its knowledge base, and iteratively upgrading its own capabilities through direct experience?

Here we present STELLA, a generalist biomedical AI agent designed around the core principle of **self-evolution** (15). STELLA learns and improves from every problem it solves, continuously enhancing its own reasoning strategies and technical abilities. Its architecture leverages four key agents—a Manager, Developer, Critic, and Tool Creation Agent—that work in concert to orchestrate complex tasks (Fig. 1A). Given a research prompt, the Manager Agent coordinates a multi-step reasoning plan. The Dev Agent then executes these steps by generating and running Python code to perform complex bioinformatics analyses. Throughout this process, the Critic Agent assesses intermediate results, identifying flaws and providing actionable feedback to refine the approach, creating a robust, iterative problem-solving loop (16–18).

The key to STELLA's advancement lies in its two novel self-evolving mechanisms. First, a **Template Library** of reasoning workflows is dynamically updated with successful strategies, allowing STELLA to learn from and generalize its problem-solving approaches over time (Fig. 1B). Second, its **Tool Ocean**, which contains STELLA's accessible bioinformatics tools, databases, and APIs, is not fixed. The Tool Creation Agent can autonomously identify, test, and integrate new tools in response to the demands of a novel problem, ensuring its capabilities are never limited to



Input: Uncover the mechanism of **acquired chemotherapy resistance** in a patient's tumor and propose a targeted **re-sensitization strategy**

Dataset Preprocessing: Receiving and preprocessing the two scRNA-seq datasets (pre-treatment and post-relapse) with normalization

Cell State Annotation: Get annotation of all cell types and states across both timepoints

Differential Analysis: Identify genes and pathways that define the resistant cancer cell state

Summarization: Consolidate results and analysis into a prelimianry report

Create Conda Environment:

```
$ conda create --env -n stella-scrna
$ conda install gseapy scanpy scgpt
```

Create and run scripts:

```
def main (args):
    data = preprocess(args.data)
    data = annotate_with_scgpt(data)
    run_differential_analysis(data)
```

>> This hypothesis is correct but not actionable. scGPT has described **what changed**, but it has not explained the **underlying regulatory logic.**

>> The current analysis is insufficient. We need to find the '**keystone**' gene maintaining the resistance network. I suggest creating a tool that can perform a **virtual screen** using single cell perturbation prediction model to predict which perturbation will best revert the resistant cells to a **drug-sensitive state**.

Final Result: Cisplatin resistance in this patient's tumor is actively maintained by a gene regulatory network controlled by the master transcription factor ***MTF1***.

Figure 1: Overall Framework of STELLA, a self-evolving LLM Agent for Biomedical Research. (**A**) STELLA leverages four key agents, including Manager Agent, Dev Agent, Critic Agent, and Tool Creation Agent. The manager agent coordinates all agents and curates a reasoning template library to leverage successful reasoning experience; dev agent focuses on environment building, code creation, model training, and report writing; critic agents reflects on the intermediary results and provide suggestions; tool creation agent identifies the gap of agent capabilities and create new tools stored in Tool Ocean. Human expert and wet experiment results can provide valuable feedback and guidance in the loop. (**B** and **C**) Two key features of STELLA's self-evolving mechanisms. The Template Library evolves by including successful previous examples; the Tool Ocean evolves from simple predefined tools during agent inference.

a predefined set (Fig. 1C). This integrated system allows STELLA to not only tackle challenging, large-scale biomedical problems with high efficiency but also to grow more capable with experience. We demonstrate that STELLA achieves state-of-the-art performance across a suite of demanding biomedical benchmarks (19, 20) (Fig. 2A) and, crucially, that its accuracy systematically improves

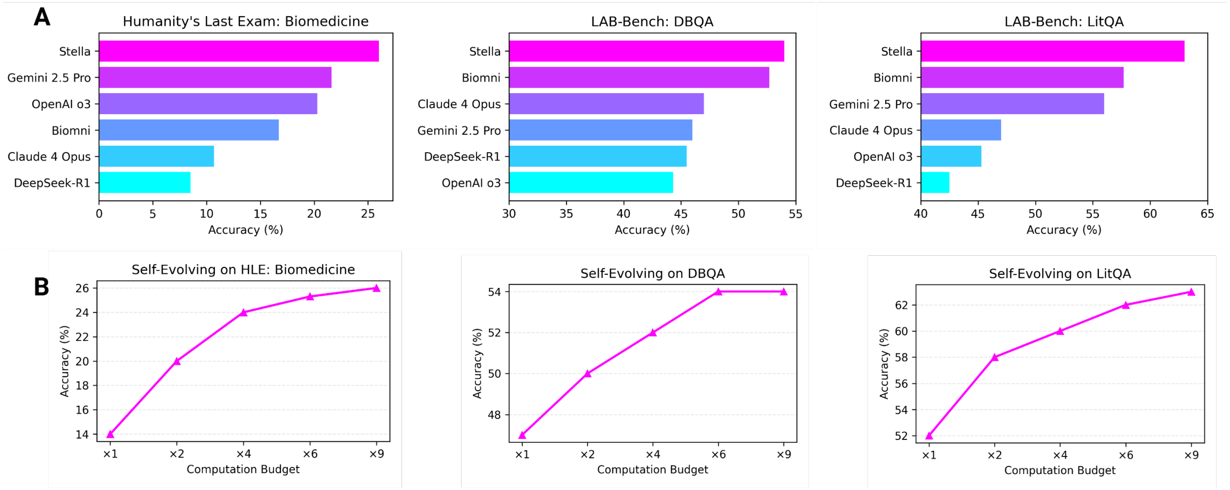


Figure 2: (**A**) Benchmark results of Stella with state-of-the-art LLMs and agents on Humanity's Last Exam: Biomedicine and LAB-Bench: DBQA and LitQA. (**B**) Test-time self-evolving effects on Benchmarks. The computation budget indicates the number of trials. The reported results represent the average accuracy across three independent evaluation runs.

with use, providing direct evidence of its self-evolving design (Fig. 2B). STELLA thus represents a significant advance towards AI systems that can learn and grow like a human scientist, dynamically scaling to meet the ever-expanding challenges of biomedical discovery.

Our empirical evaluations validate the efficacy of STELLA's architecture. Across a diverse suite of challenging biomedical reasoning benchmarks, STELLA consistently establishes a new state of the art, achieving top accuracy scores of approximately 26% on Humanity's Last Exam: Biomedicine, 54% on LAB-Bench: DBQA, and 63% on LAB-Bench: LitQA, outperforming the next-best models by up to 8 percentage points (Fig. 2A). Critically, we provide direct evidence for its core self-evolving capability. This improvement is substantial; with increased computational experience, STELLA's accuracy on HLE: Biomedicine benchmark almost doubles, rising from 14% to 26% (Fig. 2B). This validates that STELLA not only performs at a superior level but also grows more capable with experience, effectively learning how to be a better scientist over time.

Results

STELLA's Overall Framework

STELLA leverages four key agents—a **Manager Agent**, **Dev Agent**, **Critic Agent**, and **Tool Creation Agent**—to systematically address complex biomedical research questions (Figure 2A). The workflow begins when the **Manager Agent** receives a high-level research goal, such as to "uncover the mechanism of acquired chemotherapy resistance and propose a re-sensitization strategy." The Manager Agent analyzes this goal and, guided by its reasoning experience, establishes a "Reasoning Pathway"—a strategic plan that decomposes the problem into logical steps like 'Differential expression analysis' and 'Identify keystone gene'. It first assigns the initial data analysis tasks to the **Dev Agent**, which acts as a computational workhorse. The Dev Agent creates a self-contained conda environment and executes practical analysis scripts—for instance, running `diff_analysis.py` to compare the transcriptomes of pre-treatment and post-relapse tumor samples.

The results of this initial analysis are then summarized and passed to the **Critic Agent** for rigorous evaluation. In the chemoresistance example, the Critic provides crucial feedback such as: *This hypothesis is correct but not actionable... It has described* ***what*** *changed but not the underlying regulatory logic. We need to find the 'keystone' gene.* This feedback identifies a critical capability gap. In response, the Manager Agent tasks the **Tool Creation Agent** to close this gap. This agent searches existing resources and leverages a powerful collection of predefined models and tools called the **Tool Ocean**—which includes models like to build, test, and validate a new, more powerful tool, such as a virtual perturbation screening model based on virtual cell foundation model state (21). By deploying this new tool, STELLA moves beyond simple description to prediction, ultimately identifying the transcription factor `MTF1` as the keystone regulator of the resistance network.

STELLA's Self-evolving Mechanisms for Biomedical Research

A defining feature of STELLA is its dual self-evolving capability, which allows it to learn from experience and continuously expand its own abilities (Figure 1B, C). The first mechanism is the evolution of its **Template Library**. The successful multi-step workflow used to identify `MTF1`—from initial descriptive analysis to the pivot towards a predictive virtual screen—is not discarded. It is distilled into a new, high-quality reasoning template and saved in the library. This process refines STELLA's strategic knowledge, allowing it to solve similar ”mechanism of resistance” problems more efficiently in the future.

The second, more profound level of evolution is the expansion of the **Tool Ocean**, a dynamic and growing collection of STELLA's executable capabilities. This ocean contains a diverse array of computational tools that can be broadly classified into three main categories: (1) functions for querying established scientific databases, (2) interfaces for leveraging large-scale foundation models, and (3) customized analysis tools. The first category provides direct access to vital data sources like `PubMed` (22), `ClinVar` (23), and protein structures from `PDB` (24). The second allows STELLA to harness the power of state-of-the-art AI, including models like `AlphaFold 3` (25) for protein structure prediction, `scGPT` (26) for single-cell data interpretation, and `ESM3` (27) for protein language modeling. The third category consists of specialized and custom-built scripts for tasks like network analysis, and data integration.

Together, the evolution of Template Library and Tool Ocean empowers STELLA to tackle increasingly complex biomedical challenges with growing autonomy and scientific sophistication.

STELLA Outperforms State-of-the-art LLMs and Agents

To evaluate its effectiveness, STELLA was benchmarked against a suite of state-of-the-art large language models and specialized agents on three challenging biomedical question-answering tasks. The results, presented in Figure 2A, demonstrate that STELLA consistently achieves superior performance across all benchmarks. On the `Humanity's Last Exam (Biomedicine)` (28) benchmark, STELLA attained a top accuracy of 26%, surpassing all other tested models. This lead was extended on the `LAB-Bench` (20) suite, where STELLA achieved the highest scores of approximately 54% on the DBQA task and 63% on the LitQA task. These results validate the efficacy of its integrated multi-agent architecture compared to both generalist models like Gemini 2.5 Pro (29) and other specialized agents.

Furthermore, Figure 2B highlights a core strength of the framework: its test-time self-evolving capability. These results show a clear and positive correlation between computational budget and performance. As the number of iterative trials increases from 1x to 9x, STELLA's accuracy exhibits a consistent and significant improvement across all three benchmarks. For instance, on the LitQA task, accuracy rises from approximately 52% at a 1x budget to 63% at a 9x budget. This demonstrates that STELLA's self-evolving design effectively leverages increased computation to refine its strategies, correct errors, and ultimately enhance the quality of its final answer.

Conclusion

In this work, we addressed a fundamental limitation of current AI agents in biomedical research: their reliance on static, predefined capabilities in a field defined by constant evolution. We introduced STELLA, a generalist agent built on the principle of self-evolution. By integrating a multi-agent architecture with two novel mechanisms—an adaptive Template Library for reasoning and a dynamic Tool Ocean for capabilities—STELLA can learn from experience, continuously expanding its own knowledge and skills.

Our results demonstrate that this self-evolving design is not only effective but transformative. STELLA not only achieved state-of-the-art performance across multiple challenging biomedical benchmarks but, more importantly, showed significant and systematic improvement as it gained experience. This capacity to learn and grow moves beyond simple automation and represents a paradigm shift from AI as a static tool to AI as a dynamic scientific partner.

The development of STELLA marks a critical step towards creating truly autonomous AI scientists that can keep pace with the rapid rate of discovery. While challenges remain in bridging the gap between benchmark performance and real-world laboratory application, the ability of an agent to autonomously identify and master new tools lays the groundwork for systems that can explore novel scientific frontiers. Future work will focus on deploying STELLA in real-world research workflows and enhancing its collaboration with human scientists. Ultimately, self-evolving agents like STELLA have the potential to democratize expertise, unlock new avenues of inquiry, and fundamentally accelerate the engine of biomedical discovery.

Methods

Baselines

To evaluate STELLA's performance against existing methods on Humanity's Last Exam and the LAB-Bench datasets, a comprehensive set of baseline models was selected, categorized into two main groups:

- **LLMs:** We included Gemini 2.5 Pro (29), Claude 4 Opus (30), DeepSeek-R1 (31), and OpenAI o3 (32) as representative state-of-the-art LLMs that offer strong general knowledge and reasoning capabilities. Gemini 2.5 Pro is known for its large context window and strong performance on complex tasks. Claude 4 Opus is a highly capable model recognized for its advanced reasoning across a wide range of benchmarks. DeepSeek-R1 is noted for its advanced language understanding and reasoning skills, while OpenAI o3 represents a powerful, state-of-the-art model from OpenAI.
- **Biomedical Agents:** Biomni (14) was chosen as a domain-specific baseline. As a powerful agent explicitly designed to automate and advance biomedical research across a wide range of subfields, it provides the most direct and relevant comparison to STELLA's performance in this specialized domain.

For STELLA, we use Claude 4 Sonnet for the Dev Agent and Tool Creation Agent. Gemini 2.5 Pro is used for the Manager Agent and Critic Agent.

Q&A Benchmarks

For a direct and fair comparison with leading LLMs and agents, we followed the experimental settings of Biomni (14) and OriGene (13) with some modifications and applied to two main benchmark suites:

- **LAB-Bench (DBQA & LitQA):** (28) The testing sets were created by using the same 12.5% sampled subset of the complete Database Question-Answering (DBQA) and Literature Question-Answering (LitQA) sub-benchmarks. No development sets are used. This provides a cost-effective yet representative assessment of model performance. Our evaluation strictly followed the official LAB-Bench protocol, using multiple-choice answer options and allowing for abstention due to insufficient information.

- **Humanity's Last Exam (HLE):** (33) We followed the sampling protocol from the Biomni study, evaluating STELLA on a selected set of 50 representative questions from the benchmark. This question set spans fourteen subdisciplines of Biology and Medicine, including Genetics, Molecular Biology, Computational Biology, and Bioinformatics. The evaluation was conducted on the test set following the established protocol.

References and Notes

1. R. Botvinik-Nezer, *et al.*, Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582** (7810), 84–88 (2020).

2. I. Thiele, B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5** (1), 93–121 (2010).

3. E. Gibney, R. Van Noorden, Scientists losing data at a rapid rate. *Nature* (2013), doi:10.1038/nature.2013.14416.

4. H. Wang, *et al.*, Scientific discovery in the age of artificial intelligence. *Nature* **620** (7972), 47–60 (2023).

5. G. Tom, S. Li, J. M. Gregoire, A. Aspuru-Guzik, Self-driving laboratories for chemistry and materials science. *Chemical Reviews* **124** (16), 9633–9732 (2024).

6. C. Peng, *et al.*, A study of generative large language model for medical research and healthcare. *NPJ digital medicine* **6** (1), 210 (2023).

7. Y. Qu, *et al.*, Crispr-gpt: An llm agent for automated design of gene-editing experiments. *bioRxiv* pp. 2024–04 (2024).

8. K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, J. Zou, The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv* pp. 2024–11 (2024).

9. Y. Roohani, *et al.*, Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. *International Conference on Learning Representations* (2025).

10. E. Wang, *et al.*, Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196* (2025).

11. S. Gao, *et al.*, Txagent: An ai agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970* (2025).

12. Y. Xiao, *et al.*, Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *BioRxiv* pp. 2024–05 (2024).

13. Z. Zhang, *et al.*, OriGene: A Self-Evolving Virtual Disease Biologist Automating Therapeutic Target Discovery. *bioRxiv* pp. 2025–06 (2025).

14. K. Huang, *et al.*, Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv* pp. 2025–05 (2025).

15. J. Qiu, *et al.*, Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286* (2025).

16. J. Wei, *et al.*, Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022).

17. S. Yao, *et al.*, React: Synergizing reasoning and acting in language models, in *International Conference on Learning Representations (ICLR)* (2023).

18. X. Wang, *et al.*, Executable code actions elicit better llm agents, in *Forty-first International Conference on Machine Learning* (2024).

19. L. Phan, *et al.*, Humanity's last exam. *arXiv preprint arXiv:2501.14249* (2025).

20. J. M. Laurent, *et al.*, Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362* (2024).

21. A. K. Adduri, *et al.*, Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv* (2025), doi:10.1101/2025.06.26.661135v1, `https://www.biorxiv.org/content/10.1101/2025.06.26.661135v1`.

22. J. White, PubMed 2.0. *Medical reference services quarterly* **39** (4), 382–387 (2020).

23. M. J. Landrum, *et al.*, ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44** (D1), D862–D868 (2016).

24. J. L. Sussman, *et al.*, Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Biological Crystallography* **54** (6), 1078–1084 (1998).

25. J. Abramson, *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630** (8016), 493–500 (2024).

26. H. Cui, *et al.*, scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods* **21** (8), 1470–1480 (2024).

27. T. Hayes, *et al.*, Simulating 500 million years of evolution with a language model. *Science* p. eads0018 (2025).

28. L. Phan, *et al.*, Humanity's last exam. *arXiv preprint arXiv:2501.14249* (2025).

29. Google, Gemini Pro, Web page (2024), `https://deepmind.google/models/gemini/pro/`, accessed: July 1, 2025.

30. Anthropic, Introducing Claude 4, Blog post (2025), `https://www.anthropic.com/news/claude-4`, accessed: July 1, 2025. Note: This is a hypothetical URL provided for citation formatting.

31. D. Guo, *et al.*, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

32. OpenAI, Introducing o3 and o4-mini, Blog post (2025), `https://openai.com/index/introducing-o3-and-o4-mini/`, accessed: July 1, 2025. Note: This is a hypothetical URL provided for citation formatting.

33. J. M. Laurent, *et al.*, Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362* (2024).