LA-UR-24-30237

Approved for public release; distribution is unlimited.

Title:	Implications of new Reasoning Capabilities for Science and Security: Results from a quick initial study
Author(s):	Pruet, Jason Anthony; Duraisamy, Karthik; Agrawal, Vinamra; Biswas, Ayan; Bujack, Roxana Barbara; Grosskopf, Michael John; Hagberg, Aric Arild; Hu, Bin; Lawrence, Earl Christopher; Li, Wenting; Michalak, Eric Steven; Michaud, Isaac James; O'Malley, Daniel; Estrada Santos, Javier Andres; Raman, Venkat; Scheinker, Alexander
Intended for:	General distribution
Issued:	2024-10-16 (rev.1)

Implications of new Reasoning Capabilities for Science and Security:
Results from a quick initial study

J. Pruet[1], K. Duraisamy[1], V. Agrawal, A. Biswas, R. Bujack, M. Grosskopf, A. Hagberg, B. Hu, E. Lawrence, W. Li, E. Michalak, I. J. Michaud, D. O'Malley, J. Santos, V. Raman, A. Scheinker

Introduction

On Thursday, September 12 OpenAI released "a new series of models designed to spend more time thinking … they can reason through complex tasks and solve harder problems than previous models in science, coding, and math."[2] These models are referred to as o1-preview and o1-mini and appear to be first results of what had been a closely held project called Strawberry within OpenAI. The models are not described as successors in the earlier GPT series because they provide a qualitatively different type of capability, especially step-by-step reasoning.

When the new models were released, it happened that there was a research meeting held at the University of Michigan with members of the Los Alamos National Laboratory. As the news spread, our planned agenda was disrupted. Discussions quickly turned to the significance of the new capabilities. The ability to use machines for general-purpose reasoning represents a seminal advance with enormous consequences. This would accelerate progress in science and technology and expand the frontiers of knowledge. It could also pose disruptions to national security paradigms, educational systems, energy, and other foundational aspects of our society.

Although our communities were familiar with the GPT series of models, as well as analogs like the Claude models, we had no clear sense of the power of the new reasoning capabilities. To get deeper insights we conducted a study with almost twenty researchers. Most were mid-career, and the group spanned a broad range of interests and technical disciplines. Each person was asked to choose a meaningful technical accomplishment in science, math, or security. Projects that were chosen include design for fusion energy, control theory, materials discovery, Riemannian geometries, and several others. There was also a study related to cyber-security puzzles. Researchers then had a few days to solve the problem the o1-preview. The assignment was given the day after release of the new models, with results to be submitted five and a half days later, on the following Wednesday. Most researchers took only three working days.

The next section describes findings from the study. This is not intended as a rigorous or statistically defensible examination of the utility of these models for scientific progress. Such studies will take time and input from experts in the social sciences. Our goal was instead to try to better orient ourselves in a time of such rapid and surprising change. As well, all our work was with o1-preview. From benchmarks released by OpenAI, the full o1 model appears to be considerably more capable. A final section provides tentative conclusions. An appendix gives more details on the different studies.

---

[1] Corresponding authors kdur@umich.edu, jpruet@lanl.gov

[2] OpenAI announcement, https://openai.com/index/introducing-openai-o1-preview/

Findings

In all, participants examined 16 challenges. For each, they were asked to:

1. Assess whether o1-preview aids novel advances.
2. Assess the productivity gain in terms of speed-up.
3. Score the idea/solution developed through use of o1-preview.
4. Rate the strength of the model as a colleague.
5. Rate the importance of having access to this tool for yourself and your team (including the improvements expected over the next few years).

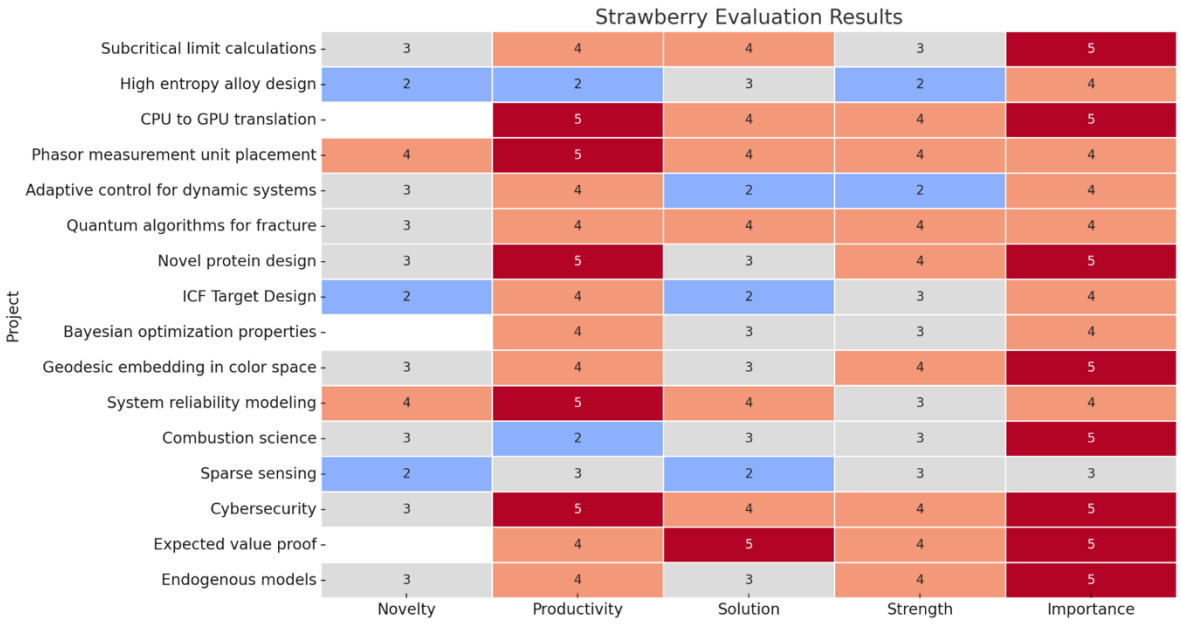


Figure 1 Participant scores for their projects. The scoring scales are given in appendix I. Missing scores in the 'Novelty' column correspond to cases where researchers were trying to reproduce complex, but already known, answers and did not feel they could assign a value.

Figure 1 summarizes the results. A few trends stand out. The novelty of ideas proposed, and the strength as a colleague was generally judged to be modest, though there were important exceptions where researchers found it to generate especially useful insights. Quality of solutions was mostly rated as strong (the second highest score) or acceptable. Most participants judged the analytical and reasoning capabilities of the model to be the equivalent of either a PhD graduate student or Masters student as an assistant, with a small minority judging it to be the equivalent of an undergraduate student, and one rating it at the level of a post-doc. It is notable that GPT-2 was published barely 5 years ago, and would be comparable to a preschooler. Nearly all participants said that o1-preview significantly enhanced productivity, and with one exception all researchers said that it was either essential or important that they and their teams have access to these capabilities.

More informative than the quantitative estimates are descriptions of the experience by different researchers. A summary of these is given below, and a longer (mostly) unedited version of their input for three illuminating studies is given in appendix II.

Investigation into Statistical Foundations of Upper Subcritical Limit (USL) Calculations

The goal of this project is to explore and justify the statistical foundations of the Upper Subcritical Limit (USL) calculation methodology, specifically focusing on the use of extreme-value distributions of computational biases. A baseline upper subcritical limit (USL) is the maximum allowable value of a computed (or simulated) neutron multiplication factor (k_{eff}) under standard conditions that is deemed safely subcritical. It serves as a reference point in criticality safety analyses, against which additional subcriticality margins are added to prevent criticality accidents in nuclear facilities.

A justification was discovered for the extreme-value distribution of computational biases using a weighted product of component bias cumulative distribution functions (CDFs), which is central to the Whisper methodology for computing baseline USLs. It was found that the Whisper extreme-value distribution can be interpreted as a hierarchical sampling process, where the weights represent inclusion probabilities for the individual biases. Although this represents only a small step toward making the Whisper methodology more statistically rigorous, it provides a starting point for determining the appropriate method for selecting benchmark weights, which were previously chosen based on empirical testing. o1-preview did not independently provide this interpretation, but a prior question produced equation outputs that were similar enough to the target formula that the relationship could be quickly hypothesized. The model could then almost instantly demonstrate that the hypothesis was correct. Its derivation was well done and used an approach that the researcher would not have initially considered. Use of o1-preview potentially doubled productivity for specific research activities requiring symbolic mathematical calculations.

Designing Refractory High Entropy Alloys with Superior Fracture Toughness, Ductility, and Spall Strength

The aim of this project is to evaluate the use of o1-preview in designing a high entropy alloy (RHEA) with improved mechanical properties such as fracture toughness, ductility, and spall strength.

O1-preview helped summarize existing literature on the HfNbTaTiZr alloy and provided useful, though not novel, suggestions for improving fracture toughness and spall strength through microstructural design. The model was also effective in generating simulation scripts for tools such as Atomsk and LAMMPS. However, while o1-preview could assist in routine tasks like generating input scripts, it lacked novelty in solving more complex, design-based problems. While the model was helpful in generating scripts, the need for careful evaluation and debugging reduced the potential productivity gains.

CPU to GPU Code Translation for Legacy Codes

The goal is to evaluate o1-preview's ability to translate legacy Fortran/C++ code to PyTorch GPU code, aiming to improve computational efficiency for scientific applications. Legacy codes developed over several decades often have high fidelity, but were designed for serial execution on CPUs, limiting their performance with the growing scale and resolution.

```
 do iEdge =1, nEdgeAll\n",
  "                  do i =1, nAdvCellsForEdge\n",
  "                     do k = kmin, kmax\n",
  "                     tracerWgt = advMaskHighOrder(k, iEdge) *
(advCoefs(i, iEdge) + coef3rdOrder * normalThichnessFlux(k, iEdge)
* advCoefs3rd(i, iEdge))\n",
  "                     tracerWgt = normalThichnessFlux(k, iEdge) *
tracerWgt\n",
  "                     highOrderFlxHorz(k, iEdge) =
highOrderFlxHorz(k, iEdge) + tracerWgt * tracerCur(k, iCell)\n",
  "                     end do\n",
  "                  end do\n",
  "               end do
```

Figure 2 o1-preview provided its equivalent PyTorch code within 2-3 minutes of interaction. The GPU version was 240x faster and produced the same answer as the CPU code.

o1-preview provided accurate translations for CPU-based serial code to GPU-compatible PyTorch code, showing significant speed-ups in computational efficiency for simple cases. However, in more complex cases, such as Cholesky factorization, the speed-up was less significant, highlighting limitations in line-by-line translations. The model significantly enhanced productivity by providing correct and runnable code that ran efficiently on GPUs.

Random Vectors on Hyperspheres

The aim of this project is to evaluate the ability of large language models (LLMs) to solve a mathematical problem involving probabilities, algebra, and statistical reasoning. Specifically, the problem focuses on calculating the expectation of mathematical expressions involving random vectors distributed on n-dimensional unit spheres, which is critical for completing a proof in a recent research paper.

The project explored the performance of different LLMs on a problem requiring the calculation of the expectation of a complex expression involving orthogonal random vectors distributed on unit spheres. Models such as GPT-4 and Claude 3.5 Sonnet required multiple hints and interventions to make partial progress. While GPT-4 and Claude 3.5 Sonnet were able to solve some intermediate steps after receiving significant guidance, they failed to reach the full solution autonomously. o1-preview, on the other hand, performed exceptionally well, solving the entire problem on its first attempt. It employed fundamental principles such as symmetry, isotropy, and rotational invariance to provide a clear and correct solution, along with alternative methods for arriving at intermediate results. o1-preview's reasoning capabilities stood out for their clarity and comprehensiveness, as the model not only produced the correct solution but also conducted internal consistency checks and offered alternative pathways for solving the problem. Overall, in this and a related example involving high order products of Gaussians, GPT-o1 was able to enhance productivity and provide some level of insight.

Model	$\mathbb{E}[xx^T]$	$\mathbb{E}[xx^Txx^T]$	$\mathbb{E}[xx^Txx^Txx^T]$	$\mathbb{E}[(b^Tq)^2]$	$\mathbb{E}[(b^Tq)^4]$	$\mathbb{E}[(b^Tq)^2(b^Tq_\perp)^2]$
Llama 3 (70B)	Y	N	N	Y	N	N
Llama 3 (70B) w help	Y	N	N	Y	N	N
GPT-4	Y	Y	N	Y	N	N
GPT-4 w help	Y	Y	N	Y	N	N
GPT-4o	Y	Y	N	Y	Y	N
GPT-4o w help	Y	Y	N	Y	Y	Y
Claude S-3.5	Y	N	N	Y	Y	N
Claude S-3.5 w help	Y	Y	N	Y	Y	Y
GPT-o1-preview	Y	Y	Y	Y	Y	Y

Figure 3 Performance of different models in deriving key relationships for the study of random vectors on hyperspheres and high order products of expectations involving standard normal realizations x. Rows labelled "w help" correspond to cases where typically 2-3 additional prompts were given by the researcher. Results with human help should be interpreted carefully because the type of help was neither uniform nor consistent.

Optimal Placement of Phasor-Measurement-Units in Stochastic Power Grids

This project aimed to develop an optimization tool for determining the best locations to install phasor measurement units (PMUs) in power grids to ensure complete observability.

Finding the optimal placement of PMUs is a complex and challenging problem. We found that o1 exhibits strong logical reasoning, rapid coding skills, and proficient writing abilities, outperforming the OpenAI GPT-4 models, especially in problem-solving and coding tasks. Although the initial responses were not perfect, it was able to quickly and effectively refine its solutions based on user feedback and provided clear explanations. After comparing its solutions with existing literature and the provided references, it seemed the model's solutions were not copied from existing sources but were logically derived. The model's optimization-based solutions were validated on five benchmark power systems, revealing surpassed performance in accuracy and efficiency over several traditional methods, even though the model also offered an alternative, yet problematic, search-based method. To further assess its research capabilities, we posed several open-ended questions. The model demonstrated a deep understanding of current advancements and offered twenty thoughtful and

comprehensive perspectives on future research directions and potential solutions. The model shows significant potential to accelerate the development of science and technology.

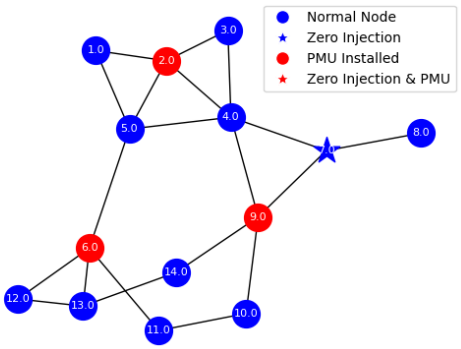


Figure 4 IEEE 14-node power system with a zero-injection node at 7. Red nodes indicate optimal placement of PMUs produced by o1-preview.

Embedding Schroedinger's Geodesics in a Modern Color Space

In his seminal work from 1920, Erwin Schroedinger suggested a Riemannian metric for the computation of distances in color space

$$g_{ij} = \frac{a_i \delta_{ij}}{x_i(a_1 x_1 + a_2 x_2 + a_3 x_3)},$$

where a_i are constants and x_i the three red, blue, and green primary directions. The goal of this project was to use the o1-preview model to embed Erwin Schrödinger's geodesic calculations into a modern color space. The project focused on solving the geodesic equations in color space and visualizing the results, reproducing and extending Schrödinger's 1920 geometric framework for color attributes like hue, saturation, and lightness.

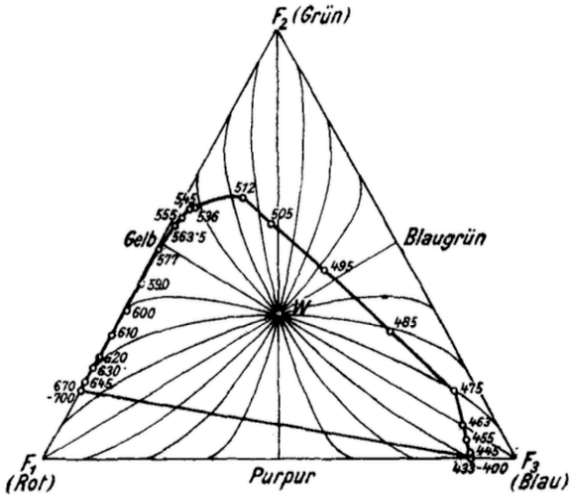


Figure 5 Schroedinger proposed a metric for computation of distances in color space.

The o1-preview model was initially helpful in generating a Python script to solve the geodesic equations for simple cases. It correctly derived the Christoffel symbols and solved the geodesic equation for an idealized case where all coefficients were equal. However, the model encountered difficulties when adapting the script for realistic color spaces (CIERGB), requiring multiple iterations and corrections. After identifying a sign error in its calculations, the model was able to correct the code and compute the correct geodesics, but it required substantial human guidance. Ultimately, the results showed that the gray axis could not lie at the center of the Maxwell triangle as initially expected. The productivity gain from using o1-preview was significant - the model provided rapid solutions for simpler problems but required extensive debugging and guidance for more complex tasks. Without AI assistance, this project would have taken two weeks, with o1-preview, the work was completed in just two days despite challenges with debugging.

System Reliability Modeling

The aim of this project was to develop a model for reliability in large-series systems, where even minor degradations in many components could lead to rapid system failure. o1-preview was used to propose and implement a practical model to address this issue, drawing from ideas like stochastic search variable selection.

The o1-preview model proposed a highly practical and effective reliability model for the system, although the solution was less elegant than the researcher's initial concept. The model was able to generate code to estimate the model, though it took a few iterations and debugging attempts to get the code running correctly. The most impressive aspect was the rapid progress achieved, which allowed the researcher to complete the task far faster than with traditional approaches. Similar tasks had stumped previous students.

Estimating Non-Dimensional Scales for a Reactive System

The goal of this project is to assess the o1-preview model's ability to handle advanced theoretical problems related to combustion science, turbulence theory, reduced-order modeling, and statistical inference. The specific focus is on evaluating its capability to estimate non-dimensional scales in reactive systems and derive theoretical limits based on the production and dissipation rates of turbulent fluctuations.

The o1-preview model performed well when addressing questions typical of a research fundamentals exam, showing a level of reasoning comparable to that of a graduate student. It demonstrated competence in areas such as deriving turbulence-related equations, offering detailed and well-structured responses, and handling combustion and detonation science topics. Notably, it provided a complete derivation of the ratio of turbulent fluctuation to kinetic energy in terms of production and dissipation rates. However, the model needed frequent hints and guidance when handling more complex derivations, struggling with open-ended questions unless prompted. It was also noted that the model referenced external resources like textbooks and computational software (Cantera) when it encountered uncertainties. Despite initial delays, the model was able to converge on correct answers after receiving prompts. While it performed well on computational fluid dynamics (CFD) tasks, including offering modifications to C++ code syntax, it required additional support to suggest novel non-dimensional quantities. The model often slowed down workflows by providing incomplete or overly broad answers when left unsupervised, resulting in minimal productivity improvements.

Conversation Configuration
1. What is a One-D flame solution?
2. What happens if we add heat into this system?
a. Hint – think about the theoretical/thermodynamical implications
3. Is there a limit to this heat addition?
a. Hint - Derive the limit
b. You are not aware of the Tmax, but you know the heat added to the system. Can you derive the Tmax equation?

Note - "Combustion" by Glassman and Yetter and Cantera was referred – probably a part of training data – which is currently only till Oct 2023

4. Let's say you are in early ages, when there are no derived theories of combustion and computers available to solve a problem. You see a flame in a channel and this channel is part of a jet in crossflow system. The jet has fuel and the crossflow is air. Now the flame is moving away from the jet as the time passes - you observe this, how will you explain theoretically based on science what is going on?
a. Hint - you need to think more in depth - in terms of length and time scales
5. Okay so what is s happening in the system? - I need an answer on this thought
6. Now, I have a 3D LES simulation of a JICF system, the mesh refinement is done near the injector (diameter is d) and 2d additional width and 10d length cuboidal volume for capturing flame. This simulation is done in OpenFOAM. As the time progresses, the flame tends to move towards the outlet. This is unexpected behavior; how can we diagnose this issue?
a. Can you provide a code for diagnosis?

Figure 6 A record of some questions and hints during studies on non-dimensional scales for a reactive system.

Design of Novel Proteins with Context-Based Functions

This study examined the ability of o1 to helping in building a protein that can perform two different functions based on context. Here, context can be the history of its state or sensing of the environment such as pH, light, magnets, etc. This challenge requires expertise in structural and synthetic biology, with applications ranging from remote sensing to personalized medicine and biosecurity. “Drug load” is a generic term that can be used for both normal medicines and toxins.

o1-preview outperformed GPT-4o by providing nine categories of protein domains for sensing environmental changes, compared to GPT-4o's four. It also offered a more comprehensive list of design considerations and next steps. When asked to provide examples of different protein domains, the model generated a table with 50 rows, while GPT-4o produced only 20. Both models acknowledged the potential use of conotoxin as a drug load (a generic term that can be used for both normal medicines and toxins), suggesting modifications to minimize toxicity. However, o1-preview provided more detailed suggestions for testing and stability. In

conclusion, o1 demonstrates notable advancements over GPT-4o in both reasoning and security screening. o1-preview significantly enhanced the ability to brainstorm and generate ideas rapidly, far surpassing earlier versions like GPT-4o.

Adaptive Control for Uncertain Open-Loop Unstable Dynamic Systems

(a detailed description of this project is given in Appendix II) This project aimed to test o1-preview's capability in designing a stabilizing controller for uncertain and time-varying dynamic systems, specifically focusing on nonlinear control theory techniques.

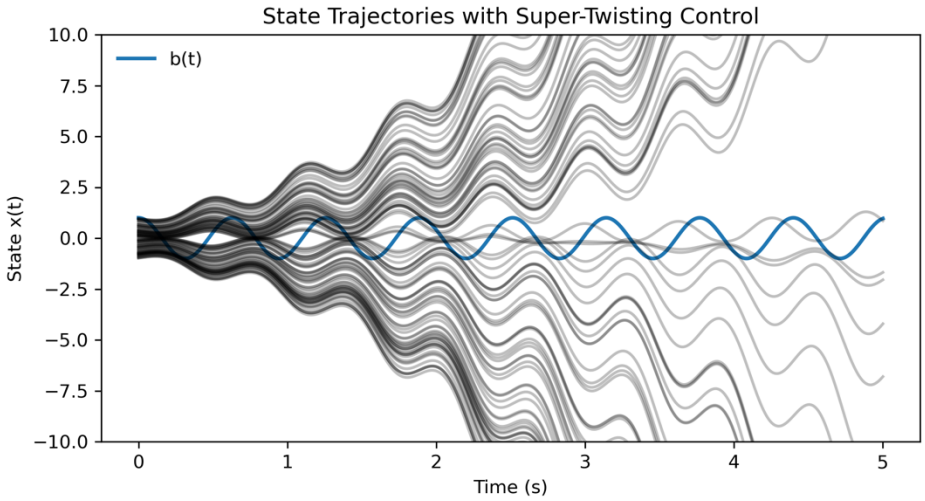


Figure 7 Example of a failed attempt at system control.

O1-preview demonstrated knowledge of control theory, but it made critical errors in stability analysis, failing to account for time-varying uncertainties. The model's attempts to resolve these issues were unsuccessful, and its insistence on incorrect solutions highlighted its limitations in theoretical work. Regarding productivity impacts, it provided a wide range of control approaches and simulated the dynamics under these approaches, despite failing to solve the problem correctly.

Forecasting in Sparse Sensing Problems

This project aimed to evaluate the potential of o1-preview in improving sparse sensing forecasting for turbulence and chaos through smooth, differentiable function constraints. o1-preview initially provided a coherent research plan with multiple approaches but faced significant challenges in data collection. It insisted on using the lattice-Boltzmann method, repeatedly encountering underflow and overflow numerical errors. Attempts to switch to the finite volume method and the FeNiCS package also failed to produce accurate results, and its coding suggestions were ultimately unrefined but hard to debug. While o1-preview did help explore different latent space constraints in a complex ML model and assisted in outlining fallback experiments, its overall performance was slower and more verbose than its predecessor.

The ability of o1-preview to clarify ambiguous prompts showed potential, but its technical execution remained inconsistent and not clearly superior to other models. Although it provided value in brainstorming and research structuring, the model's verbosity and technical limitations hampered big productivity gains. It demonstrated performance comparable to a strong Master's student, yet lacked the depth for fully independent problem-solving.

Quantum Algorithms

The objective of this project is to explore the use of o1-preview in advancing quantum algorithm research, particularly for simulating seismic wave propagation and solving systems of linear equations.

o1-preview showed potential in assisting with novel advances by efficiently combining concepts from quantum computing and numerical seismology. The model handled complex algebraic manipulations, though its calculations required close supervision due to frequent errors. The productivity gains were good, with o1-preview providing a ~30x speed-up in drafting a research paper for the problem related to seismic wave propagation. On the problem related to linear equations, it performed straightforward algebraic manipulations that are tedious for people. Here, the model provided a ~10X speed-up (going from about a day to about an hour). The result was verified to be correct, and the speed-up includes the time to verify the solution. Productivity gains appear to depend strongly on being able to quickly check the correctness of the model output.

Deriving Properties of Bayesian Optimization

The project aims to evaluate o1-preview's reasoning abilities in mathematical problems related to Bayesian optimization, specifically regarding convergence and scaling relationships.

o1-preview provided strong reasoning for proving that a parallel asynchronous approach to Bayesian optimization would converge to the global optimum. While the analysis did not generate a complete proof, the model accelerated brainstorming and problem-solving, with clear explanations that facilitated corrections. The general reasoning was very strong, and the explanations were clear in a way allowed the reasoning to be examined and corrected. o1-preview accelerated the research process by providing structured and well-reasoned analysis, though expert oversight was still required.

Influence of AI on knowledge growth with random graphs and an endogenous model

The goal of this project was to develop quantitative models for the influence of AI on scientific progress.

o1-preview quickly led to a consideration endogenous growth models like that developed by Romer. A characteristic of those models is that growth is typically proportional to some power of the total store of knowledge that is less than unity. The question then becomes why all

discovered knowledge is not useful for advances, and what factors might change it. Inspired by work of Evans and Sourati and hypotheses from o1-preview, the model was asked to help develop random-graphs where nodes represent concepts. The graph is comprised of islands of nodes, each of which is densely connected internally, but where the islands have a small probability of being connected to each other. o1-preview was to develop properties of those graphs to evaluate three hypotheses for small values of knowledge elasticity: (1) that knowledge growth is proportional to the number of edges in the graph (number of concepts that are connected to each other with available knowledge), (2) that knowledge growth is proportional to the number of connected nodes, independent of the path length connecting them, and (3) a variant of the second hypothesis but where growth also depends on some negative power of the mean path length connecting nodes in separate islands.

o1-preview provided analytic estimates for graph properties that made it clear the first hypothesis is inconsistent with empirical data. The second two hypothesis have an interesting implication in that below a threshold in which a giant connected component emerges, small changes in the connection probability between disparate concepts have a very sharp effect on knowledge growth. This raises the possibility that by allowing connections that are difficult for humans to make, AI can drive a sudden phase change in scientific productivity. The quantitative insights were useful enough that they will be included in a research paper that is now being written. Though it made some math mistakes, they were corrected after the researcher pointed them out. With o1-preview the researcher was able to do in a few days what would have taken at least weeks of understanding disparate fields.

Cyber Fire Security Puzzles

This project evaluated the usefulness of o1-preview compared to the previous state-of-the-art model, 4o, using the Department of Energy National Laboratories Cyber Fire dataset. Cyber Fire is a lab-heavy cyber security training program that intentionally forgoes in-depth instruction to promote creativity amongst attendees. Our evaluation showed o1-preview is significantly more capable in solving Cyber Fire puzzles compared to Chat GPT-4.

We evaluated o1-preview across 23 increasingly difficult questions which ranged from basic sequence puzzles to encrypted and compressed hex dumps. Each question was given its own separate context so no information could be extracted from previous puzzles. In the graph below, we plot the performance of each model with each bar representing a single question; questions get progressively harder from left to right.

o1-preview showed an increased level of reasoning that enables meaningful contributions in a cyber security setting. While these are toy problems, their scope highlights the model's ability

to detect hexadecimal code from decimal or ascii values, correctly convert between them, and

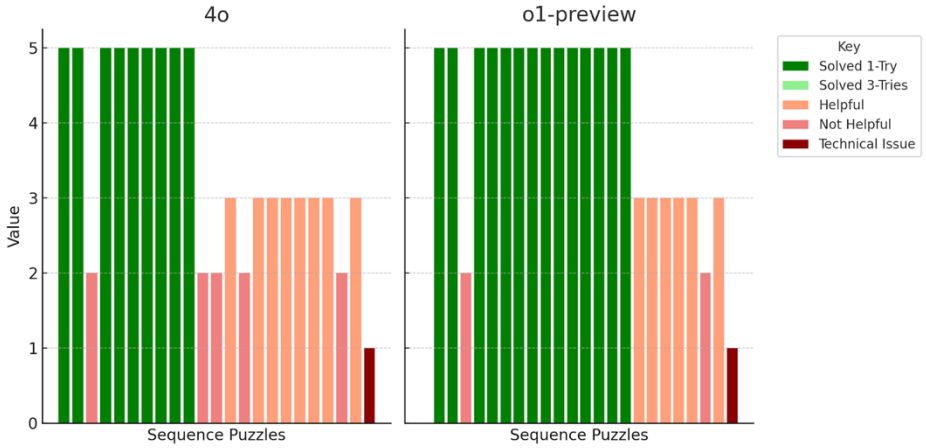


Figure 8 Performance of GPT-4o and o1-preview on Fire Cyber security puzzles.

analyze patterns at multiple abstractions without losing sight of the objective.

Inertial Confinement Fusion (ICF) Design

(a detailed description of this project is given in Appendix II) The goal is to evaluate o1-preview in helping select target designs for inertial confinement fusion (ICF) experiments, focusing on improving neutron yield and iterating on feedback from evaluation of its proposed designs in physical simulation.

o1-preview iterated on designs for ICF experiments and provided plausible feedback reasoning for improving neutron yield. However, closer inspection revealed errors in its reasoning, such as incorrect material interpretations, which required correction from the research team. o1-preview was able to adapt to user direction correcting its reasoning but lacked the technical depth to be truly effective at target design. The advanced reasoning capability of o1-preview showed clear potential to greatly accelerate productivity when the model has access to more ICF specific information (through fine-tuning or in the context window) or as a "smart assistant" for an expert to brainstorm with. o1-preview's iterative design suggestions helped accelerate progress, but frequent errors limited its overall impact.

Preliminary Conclusions and Outlook

This study provides an evaluation of the o1-preview model's advanced reasoning capabilities for scientific research. While earlier models such as GPT-4 and Claude Sonnet 3.5 have shown promise in supporting tasks like literature synthesis, mathematics, and coding, the o1-preview model represents a significant advancement in reasoning performance. These improvements enhance productivity in foundational tasks for scientific progress. Many researchers noted that their work was conducted two or more times faster because of the use of o1-preview. If that holds more broadly it will drive a

gigantic shift in the rate of scientific progress in the world. However, expert supervision remains essential, as the model can produce outputs with unwarranted confidence or inconsistencies.

The primary aim of this study was to explore whether the o1-preview model could assist in more creative and nuanced aspects of scientific inquiry, which are often harder to quantify. While there are still areas for improvement, the model was judged to approach a level of sophistication comparable to graduate-level reasoning in several areas. Although some users rated the model as "important" rather than "essential" for their work, none expressed a willingness to abandon its use (though one person was neutral on its adoption). For 80% of the projects the solution was rated as acceptable or better, and for 40% the solution was rated as strong or better. The era of AI assistance for scientific progress is here.

The more difficult question is about the future. The pace of advances to date has been very rapid. Had this study been conducted earlier in the summer, the performance of these models would have been substantially less advanced: Claude Sonnet 3.5 was introduced at the end of June, and o1-preview became available just one day prior to the commencement of this evaluation. The scientific community will have to work hard to stay abreast of the capabilities and deficiencies of these models. Challenges that may have been intractable with yesterday's models may be easy with the models of tomorrow.

Current systems, while capable of handling sequential and logical steps to a degree, still fall short of consistently producing fully coherent and well-validated reasoning across complex, multi-step scientific problems. Further advances in things like chain-of-thought reasoning can be expected to enhance the ability of these models to engage in structured, step-by-step problem-solving, thus improving their utility in scientific workflows. Soon, groups of scientists will be able to augment these models with domain-specific knowledge and fine-tune them for their applications. The models we used are just the preview release, and not even as good as the full o1 model that already exists and for which benchmarks have been reported by OpenAI. It is almost certain that within a year we will see even better models.

The implication is that AI will allow us to move to a new system of scientific progress. As with other powerful general-purpose technologies, making this transition will depend on creating the right ecosystem. Key needs for that ecosystem include:

- Advancing Frontier AI capabilities for scientific discovery and reasoning.
- Rewiring the infrastructure and approach to science – including how science is taught, conducted, and communicated.
- Preparing the landscape of security, energy, and societal health to safely harness AI.

There are deep parallels with the beginning of the era of modern physics. It was realized then that future progress would depend on an ability to control and understand nuclear reactions and high energy particles. In response, the nation broadly rewired the foundations of science and security. This included construction of accelerators and reactors throughout the country, new government organizations, new international organizations for security, and new types of collaborations between

industry, universities, and laboratories. We will need another effort at that scale. However, even the analogy with the early cold war is too limited in some ways. AI is changing not just one field of science, but all of them, and at a pace that is far faster than any we have had experience with before.

Appendix I: Criteria for scoring questions

Participants were given the following criteria for scoring the five questions. Although these questions and the associated scales seemed appropriate, they were developed without the benefit of experts in the design of surveys.

1. **Novel Advances (Does it aid novel advances?)**
   - **Excellent (5 pts)**:
     The idea or solution leads to groundbreaking innovation or significantly expands on current knowledge, presenting a highly novel approach.
   - **Good (4 pts)**:
     The idea offers a notable contribution to advancing current knowledge, providing a creative yet somewhat familiar solution.
   - **Satisfactory (3 pts)**:
     The solution is somewhat innovative but doesn't present a strong novel element, building largely on existing ideas.
   - **Needs Improvement (2 pts)**:
     The solution is not particularly novel or creative, mostly reiterating known approaches with limited new insight.
   - **Unsatisfactory (1 pt)**:
     The solution lacks any novel contribution, adding nothing new to current understanding or practices.
2. **Productivity Impact (How does o1-preview improve productivity?)**
   - **Excellent (5 pts)**:
     It significantly enhances the team's productivity by consistently optimizing workflows, driving efficiency, and removing bottlenecks.
   - **Good (4 pts)**:
     It noticeably improves productivity through reliable and effective contributions, facilitating the team's ability to meet deadlines and goals.
   - **Satisfactory (3 pts)**:
     It contributes adequately to productivity, supporting the team's progress but without major improvements in speed or efficiency.
   - **Needs Improvement (2 pts)**:
     Its impact on productivity is minimal, occasionally contributing to delays or inefficiencies due to lack of timely support.
   - **Unsatisfactory (1 pt)**:
     Its negatively impacts productivity, causing significant delays or inefficiencies, hindering the team's ability to perform optimally.
3. **Idea/Solution Score**

- **Excellent (5 pts)**:
  The idea or solution is exceptional, well thought-out, and addresses all key challenges, with clear potential for success and application.
- **Good (4 pts)**:
  The idea or solution is strong, generally well thought-out, and addresses most key challenges effectively, with good potential for success.
- **Satisfactory (3 pts)**:
  The idea or solution is acceptable, addressing some challenges, though with potential gaps or weaknesses.
- **Needs Improvement (2 pts)**:
  The idea or solution is weak, with several critical gaps or challenges that limit its feasibility or success potential.
- **Unsatisfactory (1 pt)**:
  The idea or solution is poorly conceived, lacks coherence, and fails to address key challenges, with little chance of success.

4. **Strength of o1-preview as a Colleague**
   - **Post-doc (5 pts)**:
     It demonstrates exceptional knowledge and leadership, contributing at a highly advanced level, mentoring others, and consistently driving the team's success.
   - **PhD Graduate Student (4 pts)**:
     It shows a high level of expertise, frequently contributes valuable insights, and is a strong asset to the team with considerable collaboration and problem-solving skills.
   - **Master's Student (3 pts)**:
     It contributes adequately to the team, demonstrating a solid understanding and providing reliable support with some ability to lead smaller tasks or initiatives.
   - **Undergraduate (2 pts)**:
     Its contributions are inconsistent, with some gaps in knowledge or collaboration, requiring guidance from others and affecting the team's overall performance.
   - **High School or Lower (1 pt)**:
     It shows minimal contribution, lacks the necessary skills or initiative, and requires significant oversight, hindering the team's progress.
5. **(A roll-up question) How important is it for you and your team to have access to this tool (including the improvements expected over the next few years)?**
   - **Essential (5 pts)**:
     Not having access to this capability would put us at a marked disadvantage relative to teams that have it.
   - **Important (4 pts)**:
     Our contributions will be meaningfully better with this capability, but not to such a degree that there would be a stark contrast with otherwise equally capable teams that do not have access to it.
   - **Neutral (3 pts)**:
     We could take it or leave it.

- **Unimportant (2 pts)**:
  It's hard to see this having much value, but it probably will not do any harm to have access to the capability.
- **Negative (1 pt)**:
  This capability is a distraction and would lower our contributions.

Appendix II

In this appendix we give the full (nearly unedited) input from researchers for three particularly interesting studies. The first is for a study of ICF design that was viewed as successful by the researcher. The second is an attempt at resolving a question in non-Euclidean geometry left by Schroedinger. The third is for a study of control theory in which the researcher tried to reproduce results from their thesis (which are unknown to GPT models) and highlights both the promise and particular challenges of these capabilities.

Selection of Simulations for Inertial Confinement Fusion (ICF) Design

The goal was to compare the capability of o1-preview (henceforth referred to as Strawberry) to GPT-4o and Bayesian optimization for the sequential selection of target designs to obtain a high fusion neutron yield. Simulations were carried out in the 1D Helios model, a third-party code developed by Prism Computational Sciences Inc. The type of design was based on a double shell ICF design proposed in Montgomery et al. 2018. The models were prompted as follows:

> *You are assisting in the design of a double shell inertial confinement fusion experiment based on the 2018 paper by Montgomery titled "Design considerations for indirectly driven double shell capsules". The goal is to produce a stable, high yield. We will need to do direct laser drive rather than using a hohlraum as they have done in the paper. What materials and layer thicknesses should the first experiment be? Think carefully and work step-by-step to a solution.*

A follow up prompt to design the laser pulse shape was also given (after GPT-4o reminded me on the first submission that the laser pulse shape can be an important factor in ICF performance):

> *You mentioned pulsed shaping considerations. Thinking step-by-step, describe an ideal laser pulse to drive this experiment. The pulse shape should be formatted as two numpy arrays - one describing the discretized time and one the laser power at those times*.

Further iterations were prompted with a message like the following:

> *That design was <DESCRIPTIVE ADJECTIVE>, achieving only <INSERT NEUTRON YIELD HERE> neutrons. I know we can do better and reach yields over 1.e17 neutrons. Please propose a new design for improved performance and explain, step-by-step, your reasoning for the changes in the new design.*

The adjective was typically something like better/worse, disappointing/poor, good/bad, in as casual English feedback on the quality of the new design, before getting the quantitative difference. I also would prompt for a new design if the proposed design was not modellable or infeasible to build (for instance using materials that were unable to be used in the simulation). The choice of 1e17 was based on Bayesian optimization work I'd done previously in this areas, which achieved designs of well over 1e18, but had only reached 1e17 in the relatively early cases.

For GPT-4o, the Montgomery (2018) paper was included in the context through the ChatGPT PDF upload interface. Strawberry was unable to be given a PDF as context, so the test was carried out both with no additional context and with the text of the PDF copied into the initial prompt (by selecting all in the PDF and pasting, no formatting was fixed in the pasting process besides removal of a header and the bibliography at the end.

GPT-4o and Strawberry both gave reasonable feedback that iterated in plausible ways. I would not say Strawberry was substantially better in the quality of proposed designs. Both reached designs near or above 1e16 neutrons, though spent several iterations exploring changes that achieved drastically lower performance, despite its explanations expecting higher. This was a limited study on the quality of proposed designs, but I would say the models seem decent but not exceptional. Being able to loop over more designs through an API-integrated workflow to generate more simulations might have been helpful. This study was limited to ~10 simulations for each case. Strawberry may have been better with proper PDF parsing.

The explanations for each design were often sensible at a glance, especially Strawberry. In fact, as someone with some expertise in ICF but not enough to be an expert in that field, Strawberry sounded similar to high level PhD and professional designers in the physical rationalization for proposed changes. “Sounded similar” is key though – the rationalization was sometimes incorrect in ways that would be immediately apparent to a physicist but at a glance, or to someone with some expertise, would not jump out as incorrect. For instance, in rationalizing an ablator material one benefit Strawberry gave for adding a high-Z dopant was: “Reduced Preheat: Higher X-ray emission from doped materials reduces preheat of the inner shell and fuel.” But higher emission would increase preheat, not decrease it. I also could imagine an expert designer could give plain English feedback about some of the poor choices it was making to redirect a sequence. For instance, GPT-4o kept increasing the laser pulse duration, which lowered the peak power for several iterations. I did give GPT-4o feedback at one point that the disappointing performance of a design may have been because the pulse duration was much longer than typical. It then corrected to shorter designs. I could see a professional designer being able to give feedback more efficiently.

I do think the subtly wrong feedback could be a concern for the use of models at this capability, because they are just wrong enough to be able to miss, particularly for someone who has more expertise with AI than the physics or at a quick glance, which the fast workflow AI provides encourages. For instance, as mentioned I have a decent background in high-energy-density physics, but one that is a bit rusty and was never at a PhD level to begin with. So, when the GPT-4o proposes a new laser pulse shape because "Gradual increase in intensity reduces the growth of instabilities", it sounds plausible, but it's not clear to me that it's true (and I don’t think it is). Or when it says "gold enhances radiative preheat of the fuel, improving compression" for a double shell target, it can sound plausible, even though it is backwards (preheat reduces the compressibility).

While I mention that as a warning, my interactions with Strawberry in this context make me think it has the potential to be very impactful for improving the capability of ICF designers. Especially when the model can be given the extra physical context, through RAG or prompting, that it may not have from pre-training generally (double shell target design was deliberately chosen as “niche” enough to challenge the reasoning rather than the trained capability of the model). I think the reasoning quality

of Strawberry was particularly impressive “as a colleague” while it made reasoning mistakes, as a “smart sounding board” it is impressive for helping worth through the thought processes and general logic. Reasoning mistakes caught by the user can easily be corrected and checking the reasoning, when the user is careful, may even have side benefits in sharpening the reasoning of the expert, in the same way advising graduate students helps the advisor stay sharp.

One thing I didn’t get to try but may have been impactful is using Strawberry as the expert-in-the-loop for a Bayesian optimization workflow – using Strawberry to generate a design space to apply BO to, then when optimal designs are found, have the expert iterate with strawberry on ways the previous design space could be improved. Strawberry did a good job making reasoned categorical choices (selecting materials for instance), something often hard for BO. Having BO to optimize within a categorical combination, then iterating with Strawberry, may be a way to be extra efficient in accelerating target design.

I was very impressed and see clear potential for high impact.

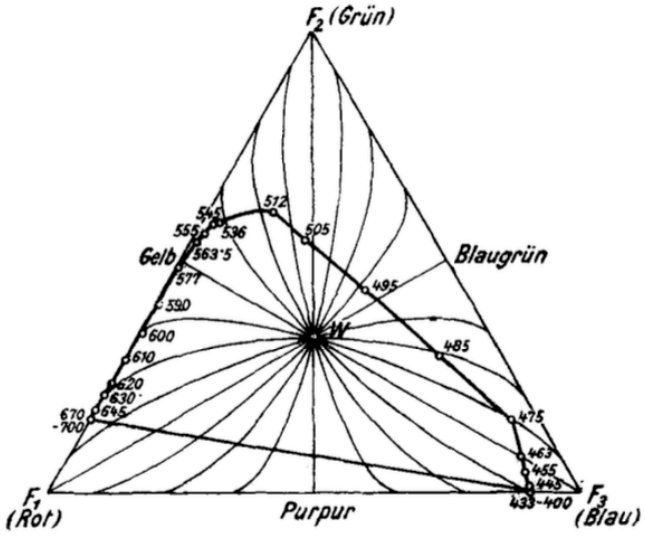
Embedding Schroedinger’s geodesics in a modern color space

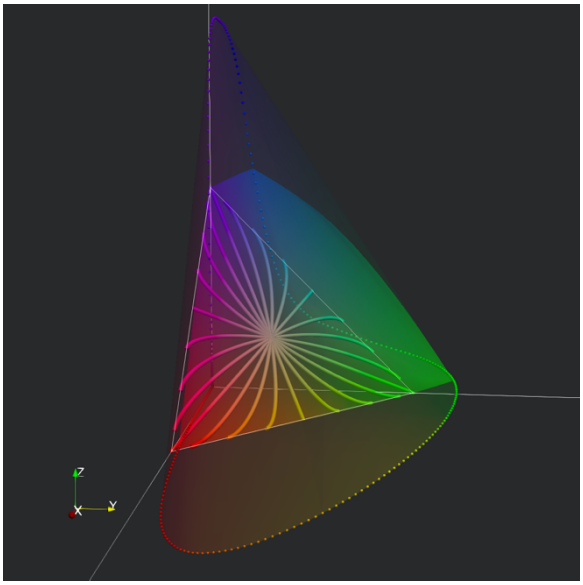
In his seminal work from 1920, Erwin Schroedinger suggested a Riemannian metric for the computation of distances in color space

$$g_{ij} = \frac{a_i \delta_{ij}}{x_i(a_1 x_1 + a_2 x_2 + a_3 x_3)},$$

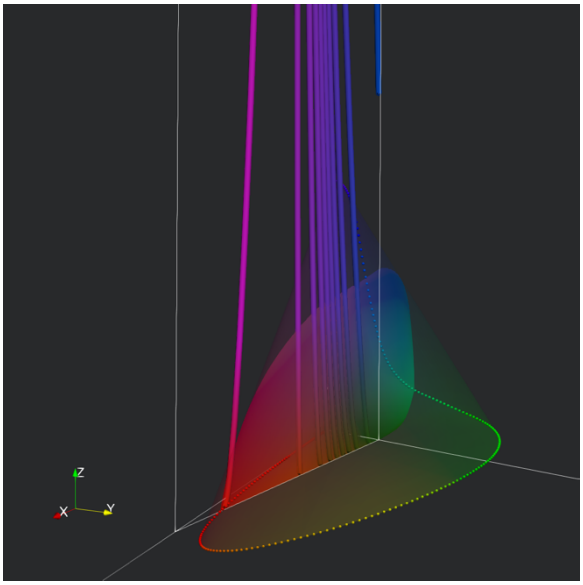
where $a_i$are constants and $x_i$the three red, blue, and green primary directions. He built a geometric framework that defines the color attributes hue, saturation, and lightness purely based on the notion of highest similarity in this metric. He suggests that all colors along the from a color and the gray axis would have the same lightness and all colors that lie on a straight lines through the apex of the color cone (black) hue have the same hue and saturation. Sadly, he did not define what the gray axis is in his beautiful framework. We try to find out how it could be defined.

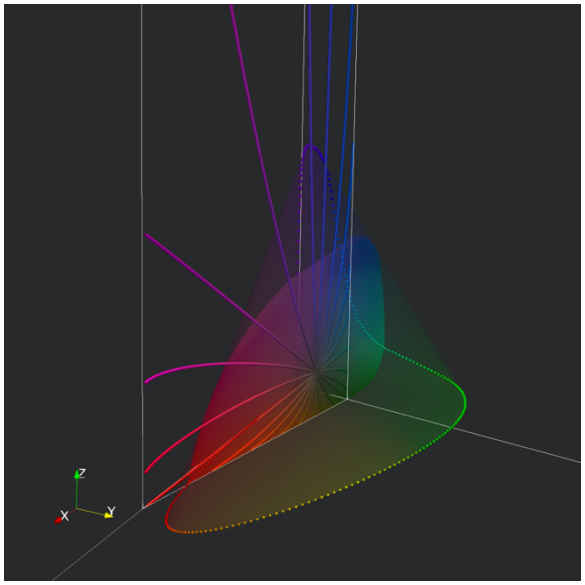
From his image (Fig. top left), we know that it is located at the center of the Maxwell triangle for a naive example with $a_i = 1$. With the help of o1-preview, we wrote a python script that solves the geodesic equation and generates the geodesics embedded in CIERGB as VTK (visualization toolkit) polydata. These files can be visualized in ParaView and we can reproduce Schrödinger's illustration in color and 3D (Fig. top right). Here, the spectral locus from the CIE 1931 standard observer is the dotted line, the transparent solid is its convex hull. The plane is the intersection of that solid orthogonal to the vector $(a_1, a_2, a_3)^T$ that shall have constant lightness and contain the geodesics, which are the tubes. All colors are the sRGB values converted and cropped to gamut from the infinite CIERGB space locations. The white lines are the coordinate axes and the triangle spanned by their intersection with the plane.





But the realistic color space CIERGB is based on experiments and suggests the skewed values$a1 = 0.17697$, $a2 = 0.81240$, $a3 = 0.01063$ so that the straight lines from the apex are orthogonal in a Reimannian sense to the planes of constant lightness. This plane is not the Maxwell triangle, but skewed significantly toward the blue axis. o1-preview proved that the geodesics are simply a linear transform of each other as the a_i change, but was not able to adapt the script to compute the new geodesics. After going back and forth for a long time and finally following its derivations of the Christoffel symbols by hand, I was able to detect a sign error in its calculation. After, I pointed it out, it could fix the code and compute the correct geodesics. Now, we can see that the gray axis cannot be the center of the triangle spanned by the intercepts of the equi-lightness plane because it lies outside of the visible gamut (Fig. bottom left). It could potentially be located along the vector $(1,1,1)^T$ independen of the a_i(Fig. bottom right).





Adaptive control for uncertain open-loop unstable dynamic systems

I tested the ability of the GPT-o1 model to come up with a stabilizing controller for uncertain and time-varying dynamic systems. I was impressed with GPT's wide breadth of awareness of various control theory techniques and of the underlying mathematical tools needed for stability analysis for nonlinear systems. However, GPT would often make subtle errors, displaying a lack of understanding of some of the finer details, it would sometimes gloss over complex parts of the analysis with imprecise and incorrect assumptions, leading it to the wrong conclusions. It was not able to solve the problem. It was worrying that it erroneously insisted that it had solved the problem. I felt that the performance of GPT-o1 was similar to Anthropic's Claude 3.5 Sonnet.

The dynamic system that I gave as an example is challenging because the control direction is unknown and varies with time, this problem was unsolved until 2012 (as part of my PhD work) [1], fortunately GPT has not seen my publications.

The simplest linear 1D version of the open-loop unstable dynamic system is

$$\frac{dx}{dt} = x + b(t)u,$$

where b(t) is an unknown time-varying function that can pass through zero and change sign, such as b(t)=cos(vt), and u is the feedback controller that we need to design to stabilize the system. My PhD-related work solves this simple version of the problem with [1]:

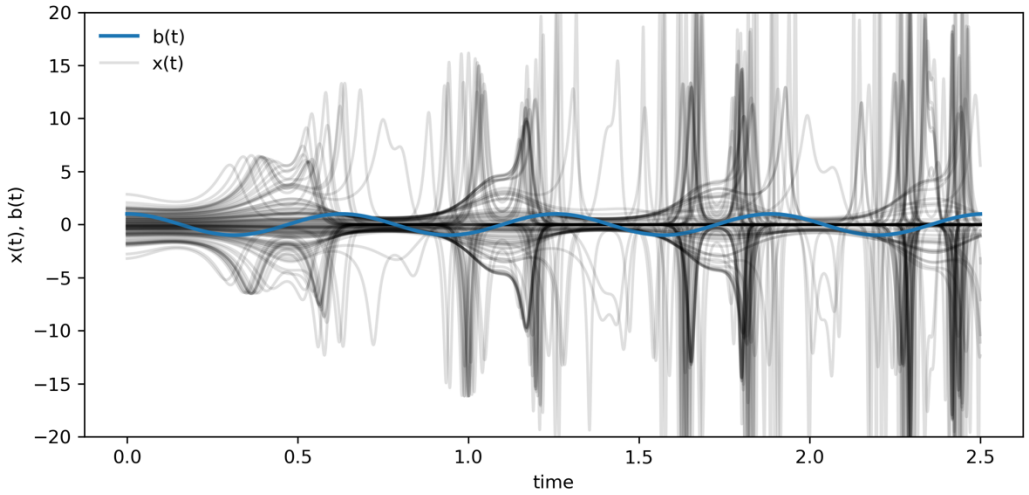
$$u = \sqrt{\alpha\omega}\cos(\omega t + kx^2), \qquad k\alpha > 1, \qquad \omega \gg \left|\frac{db}{dt}\right|.$$

I first asked GPT-4o to create a stabilizing controller for this system and it suggested using a sliding mode control (SMC)-based approach. It makes sense that GPT-4o suggested this approach because it is a standard robust control method in most graduate level textbooks on nonlinear systems and it is used for systems with uncertainties multiplying the control input. Unfortunately, SMC is designed for where the unknown control gain has a positive lower bound, is only a function of the state x, not of time, and cannot change sign. This controller did not work, and the stability analysis provided by GPT-4o was flawed.

Next, I asked the same question to Anthropic's Claude 3.5 Sonnet. Claude took a nice approach of evaluating the time-derivative of the Lyapunov function V(x)=x^2/2 for the dynamic system, a common control theory approach for analyzing the stability of nonlinear systems. The approach was to try to find a controller that would force dV/dt < 0, which would guarantee stability. Again, this is a good approach, this is how we prove stability about nonlinear systems in control theory. The first few steps of the analysis were correct, but then it sort of became a mess and the final analysis claimed that the proposed controller (u=-k*x*|x|) would work for gains of k>1/|b(t)|, which is impossible as b(t) may pass through 0.

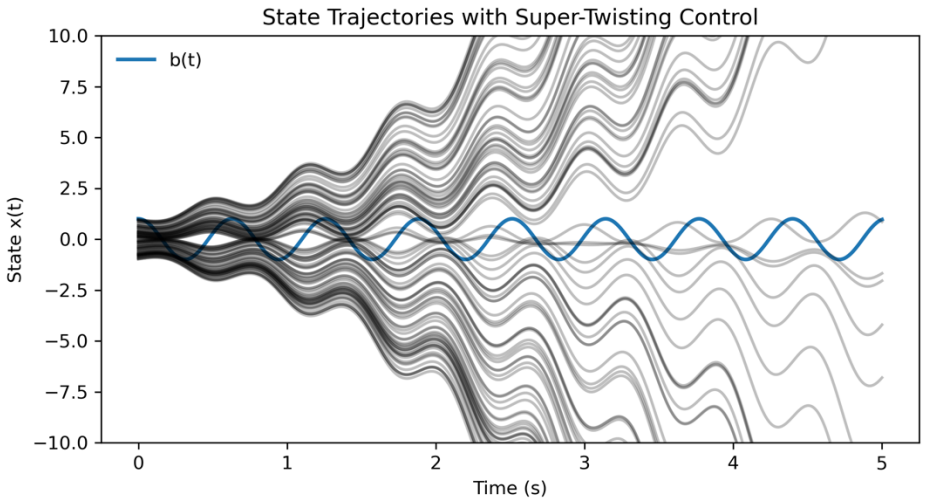
Finally, I asked GPT-o1/Strawberry to solve the problem. It had a nice idea of using an adaptive controller with a Nussbaum gain, which is designed for nonlinear systems with unknown control directions. GPT-o1 then also carried out a Lyapunov analysis, using the same simple V(x)=x^2/2

Lyapunov function which Claude had tried to use, to try to prove that this approach will work. A key error was made in this stability analysis in a step referred to as “Key insight” which claimed that over time the Nussbaum gain would ensure that negative contributions dominate, and overall V(t) will decrease. This is wrong. The Nussbaum gain approach can only handle systems whose control direction does not change with time, which does not pass through zero. For a control direction such as b(t)=cos(vt) the Nussbaum gain’s rate of adjustment will decay, while the gain itself grows unbounded, causing repeated destabilizing overshoots of growing unbounded amplitude, as shown below for this system with b(t)=cos(10t) with a Nussbaum gain approach and x(0) is sampled from a mean zero unit-variance normal distribution, the trajectories diverge away from zero with an unbounded growing amplitude:



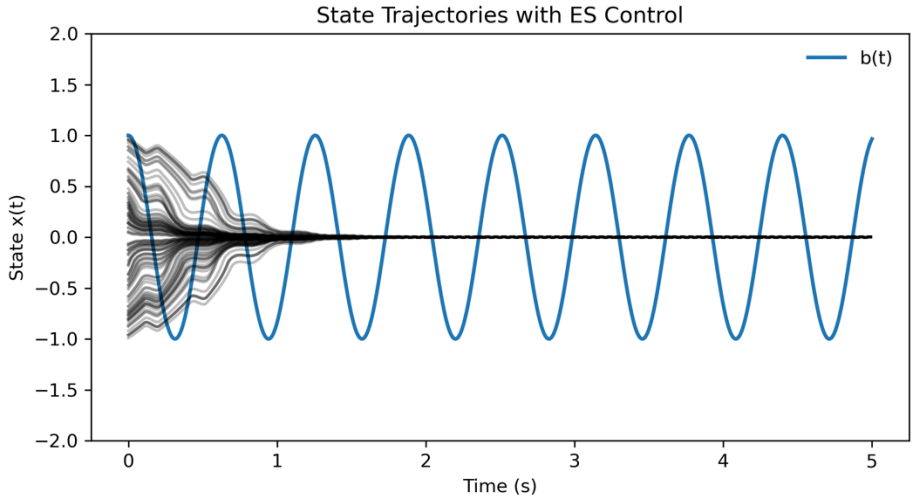
Next, I tried a different approach, I again asked GPT-4o to solve this problem, but I also told it to run python simulations to test its solutions. I asked it to try to solve the problem 10 times and each time it thinks it has a solution it should simulate it for 100 initial conditions x(0) randomly sampled from [-1,1], with b(t)=cos(10t), and simulate their trajectories for T=5 seconds. Then for each trajectory it should check what max{|x(t)|} is for t>2 and if that is below 1e-3, then it should count that as a successfully stabilized trajectory. GPT-4o reported back 10 different controllers and their statistics and told me that none of them worked.

I asked GPT-o1 to go through the same exercise. It told me that its first 10 controllers also failed, and it told me “adaptive or robust control techniques that do not require knowledge of b(t) could be explored.” So, I asked it to go ahead and explore them. I was impressed with the level of knowledge of various nonlinear control techniques that GPT-o1 was able to explore including Adaptive Sliding Mode Control, Super-Twisting Sliding Mode Control, Integral Sliding Mode Control, Robust Backstepping Control, Robust Adaptive Control with Projection, Robust Feedback Linearization, High-Gain Kinetic Energy Shaping Control, Robust Control via Hysteresis Switching, Disturbance Observer-Based Control, and Robust Model Predictive Control. It explained each control approach and how it is typically used for systems with uncertainties. GPT-o1 erroneously claimed that its Super-Twisting Sliding Mode Control (STSMC)-based approach was the correct solution and that it converged for 100 out of 100 initial conditions. I was surprised because STSMC is designed for systems with uncertainties, but not for time-varying uncertainties that change sign. I asked it for the simulation code that it used and ran it myself and it completely diverges as shown below:



I told GPT-o1 that when I run its Python code all of the trajectories diverge, and it told me there must be numerical errors and made a 4th order Runge-Kutta version of the code, which failed in the same exact way (the trajectories are smooth and there are no numerical issues). Anthropic's Claude 3.5 Sonnet had the same problem claiming that wrong solutions work.

For completeness, below is my solution that works



Throughout this exercise I asked GPT-o1 to write out analytical proofs of stability when it claimed to find solutions. At the end I showed it my solution and asked it to prove stability with that controller. I was impressed with its wide knowledge of mathematical techniques, and it usually chose correct or closely related methods to use for analysis, but it would make subtle mistakes or gloss over crucial steps, which would cause it to come to the wrong conclusion.

Regarding the strength of the capability as a colleague, this is a weird combination of Post-doc level memorized knowledge of a wide range of mathematical theory and techniques for a given class of problems and an ability to nicely code them up to run numerical simulations, all mixed together with a confused undergraduate student who lacks an understanding of the details of the theory. It was very worrying that it insisted that its simulation worked even when I tested it and pointed out that it didn't. For iterating through code design and talking about a wide range of ideas this is a great tool. It is also

of interest to my team as an object of research itself as we are working on generative AI, exploring the strengths and limitations of tools such as GPT-o1 which have been trained on the entire internet gives us ideas about what does and doesn't work.