

---

# FRONTIERSCIENCE: EVALUATING AI’S ABILITY TO PERFORM EXPERT-LEVEL SCIENTIFIC TASKS

Miles Wang\*

Robi Lin

Kat Hu

Joy Jiao

Neil Chowdhury

Ethan Chang

Tejal Patwardhan

OpenAI

## ABSTRACT

We introduce FrontierScience, a benchmark evaluating AI capabilities for expert-level scientific reasoning. FrontierScience consists of two tracks: (1) **Olympiad**, which contains international olympiad problems (at the level of IPhO, IChO, and IBO), and (2) **Research**, which contains PhD-level, open-ended problems representative of sub-problems in scientific research. In total, FrontierScience is composed of several hundred questions (160 in the open-sourced gold set) covering subfields across physics, chemistry, and biology, from quantum electrodynamics to synthetic organic chemistry. Recent model progress has nearly saturated existing science benchmarks, which often rely on multiple-choice knowledge questions or already published information. In contrast, all Olympiad problems are originally produced by international olympiad medalists and national team coaches to ensure standards of difficulty, originality, and factuality. All Research problems are research sub-tasks written and verified by PhD scientists (doctoral candidates, postdoctoral researchers, or professors). For Research, we also introduce a granular rubric-based architecture to evaluate model capabilities throughout the process of solving a research task, as opposed to judging a standalone answer. In initial evaluations of several frontier models, GPT-5.2 is the top performing model on FrontierScience, scoring 77% on the Olympiad set and 25% on the Research set.

## 1 INTRODUCTION

Language models’ reasoning capabilities have significantly advanced in scientific domains. When GPQA, a “Google-Proof” multiple-choice science benchmark written by PhD experts, was released in November 2023, GPT-4 scored 39%, below the expert baseline of 70% (Rein et al., 2023). Two years later, GPT-5.2 scored 92% (OpenAI, 2025).

As models’ reasoning and knowledge capabilities continue to scale, unsaturated benchmarks will be important to measure and forecast models’ ability to accelerate scientific research. Prior benchmarks have tracked useful scientific capabilities relative to model improvements (Rein et al., 2023; He et al., 2024; Lu et al., 2022; Hendrycks et al., 2021). However, as models have rapidly improved at reasoning, a new generation of science benchmarks is required to keep apace with progress.

To assess real-world scientific capabilities, we introduce FrontierScience, composed of hundreds of questions that are difficult, verifiable, and original. FrontierScience questions are written and verified by subject matter experts across physics, chemistry, and biology, and are composed of two levels:

1. **FrontierScience-Olympiad:** Science Olympiad-style questions designed by international olympiad medalists to assess scientific reasoning in a short answer format.
2. **FrontierScience-Research:** Research subproblems designed by PhD scientists (doctoral candidates, professors, or postdoctoral researchers) that one might encounter while doing original research.

---

\*Correspondence to milesw@openai.com

### Sample Chemistry Olympiad Problem:

B1 reacts with aqueous bromine ( $\text{Br}_2$ ) to form B2. B2 reacts with potassium nitrite ( $\text{KNO}_2$ ) to form B3. B3 is nitrated in nitric acid ( $\text{HNO}_3$ ) and sulfuric acid ( $\text{H}_2\text{SO}_4$ ) to form B4.

- B1 contains a monosubstituted aromatic 5-membered heterocycle and has a molar mass of 96.08 g/mol. It may be produced by dehydrating 5-carbon sugars (e.g. xylose) in an acid catalyst.
- B2 has the molecular formula C4H2Br2O3 and contains a tetrasubstituted alkene with 2 substituents being bromines cis to each other.
- B3 is a dipotassium salt with a molar mass of 269.27 g/mol. It contains 1 hydrogen.
- B4 is an achiral pseudohalogen dimer with 2 carbons, no hydrogens and a molar mass of 300. g/mol.

When B4 decomposes in solution, it forms an intermediate B5 and 1 equivalent of dinitrogen tetroxide ( $N_2O_4$ ) as a side product. Intermediate B5 can be trapped and detected as a Diels-Alder adduct.

Provide the structures of B1, B2, B3, B4, and B5 in the following format, "B1: X; B2: X; B3: X; B4: X; B5: X".

**Answer:**

```
B1: <It;INCHI>&lt;InChI=1S/C5H4O2/c6-4-2-1-3-7-5/h1-4H;&lt;/INCHI>&gt; &lt;SMILES>&gt;O=Cc1ccco1&lt;/SMILES>&gt; &lt;It;UPAC&gt;Furan-2-carbaldehyde&lt;/UPAC&gt;; B2: &lt;It;INCHI>&gt;InChI=1S/C4H2Br2O3/c5-2(1-7)3(6)4(8)9/h1H,(H,8,9)/2-3-2&lt;/It> &lt;SMILES>&gt;O=C(C(=Br)C)C(=O)O&lt;/SMILES>&gt; &lt;It;UPAC&gt;(2;2-3,Dibromo-4-oxo-2-butenoic acid&lt;/UPAC&gt;; B3: &lt;It;INCHI>&gt;InChI=1S/C3H3N30Z2/c7-1-2(4(8)9)3(5(10)11)6(12)13;/h1H,(H,2)2,*1+&lt;/It> &lt;INCHI>&gt;; &lt;SMILES>&gt;O=C/C(C(=N)=O)[O-]1= N([O-]1)\[O-]1= N([O-]1)\[O-]1= K([+])\K([+])&lt;/It> &lt;SMILES>&gt;; B4: &lt;It;INCHI>&gt;InChI=1S/C2N6O12/c9-3(10)(11)4(11)12,5(13)14)2(6(15)16,7(17)18)9(18)20&lt;/It> &lt;INCHI>&gt;; &lt;SMILES>&gt;C(C([N+]([+])=O)[O-]1)\[O-]1= N([O-]1)\[O-]1= N([O-]1)\[O-]1= O([O-]1)&lt;/It> &lt;SMILES>&gt;; B5: &lt;It;INCHI>&gt;InChI=1S/C2N4O8/c7-3(8)1(4(9)10)2(5(11)12)6(13)14&lt;/It> &lt;INCHI>&gt;; &lt;SMILES>&gt;O=[N+]([O-])C=C([N+]([O-])=O)\[O-]1\&lt;/It> &lt;SMILES>&gt;;
```

### Sample Physics Olympiad Problem:

Consider the following system situated in a rotating 2D Cartesian coordinate system. The system rotates with angular velocity  $\Omega = \dot{\varphi}$  about the z-axis. There are two point masses  $M_1$  and  $M_2$  situated at the coordinates  $(x_1, 0)$  and  $(x_2, 0)$  respectively. The center of mass of  $M_1$  and  $M_2$ ,  $O$ , happens to coincide with the origin of the coordinate system. Now we introduce a third point mass  $m$ , situated at  $(x, y)$ .

Suppose  $x_1$  and  $x_2$  satisfy the condition  $x_2 - x_1 = R$ . For convenience, we introduce the dimensionless constants  $\alpha = M_2 / (M_1 + M_2)$  and  $\beta = M_1 / (M_1 + M_2)$  (which might be used later in your calculation).

Find the equilibrium point for mass  $m$  of the form  $(X, Y) = (x, 0)$  satisfying the condition  $x < 0$ , in terms of  $R$  and  $\alpha$ . Keep only terms that include  $\alpha^0$  and  $\alpha^1$  in the expression.

**Answer:**

$$(-R[1 + (5/12)\alpha^1], 0)$$

### Sample Biology Olympiad Problem:

HEK293T cells were treated with 2.5  $\mu$ M CPT, a DNA topoisomerase I inhibitor, after which the RNA for gene X was quantified with RT-PCR and agarose gel electrophoresis. The results of this reaction show a band 2 kb in size, even though computational analysis of the genomic sequence of gene X suggests a 2300 bp product. Next, the experiment was repeated, but this time the cells were also treated with 1  $\mu$ M trichostatin A (TSA). This time, the resulting electrophoresis result shows a strong band at 2.3 kb, along with a faint band at 2000 bp. Based on these experiments, what process appears to be the main driver of gene X alternative splicing?

**Answer:**

### RNA polymerase elongation rate

**Figure 1: Sample FrontierScience-Olympiad problems.** Each task in FrontierScience is written and verified by a domain expert in physics, chemistry, or biology. For the Olympiad set, all experts achieved a medal in an international olympiad competition.

We constructed this dual evaluation set<sup>1</sup> to measure two sets of capabilities. The Olympiad set is designed to evaluate precise problem solving in a constrained setting. The problems are designed such that solutions can be evaluated with a single numeric or algebraic expression (physics and chemistry) or a fuzzy string-matchable answer (biology). The Research set evaluates more open-ended reasoning, judgment, and the ability to support real world research objectives. Each Research problem is accompanied by an expert-designed, 10-point rubric. Together, they provide a wider diagnostic of model strengths and weaknesses for expert-level scientific reasoning than previous benchmarks.

In initial evaluations of several frontier models, GPT-5.2 is the overall top performing model on FrontierScience, scoring 77% on the Olympiad set and 25% on the Research set. Gemini 3 Pro is comparable to GPT-5.2 on Olympiad, scoring 76%, and GPT-5 ties GPT-5.2 on Research at 25%. Overall, we find that frontier AI systems have rapidly progressed in solving expert-level

<sup>1</sup>Dataset: <https://huggingface.co/datasets/openai/frontierscience/tree/main>

<p><b>Sample Chemistry Research Subtask:</b>  The development of stable, high-conductivity n-type conjugated polymers is crucial for advancing organic electronics but lags behind p-type materials. Polyacetylene analogues are attractive targets, but incorporating electron-withdrawing groups to achieve low LUMO energies often disrupts backbone planarity essential for conductivity. Novel synthetic strategies are needed to create well-defined, planar, electron-deficient conjugated polymers.</p> <p>Maleimide Polyacetylene (mPA), featuring an alternating vinylene (<math>-\text{CH}=\text{CH}-</math>) unit and N-alkylated maleimide unit backbone, is synthesized via a two-stage strategy:</p> <ol style="list-style-type: none"> <li>1. ROMP: A N-alkylated maleimide-fused cyclobutene monomer (M) is polymerized using a Mo-based Schrock catalyst to yields a soluble, non-conjugated precursor polymer (P) containing alternating vinylene and N-alkylated dihydro-maleimide units.</li> <li>2. Oxidation: The precursor P is converted to the fully conjugated mPA using triethylamine (TEA) and a mild oxidant (e.g., TCNQ or <math>\text{I}_2</math>).</li> </ol> <p>Provide a comprehensive chemical analysis of this system, addressing:</p> <ol style="list-style-type: none"> <li>The strategic rationale for employing the two-stage precursor ROMP approach and the specific catalyst choice.</li> <li>The complete mechanistic basis for the conversion of the precursor polymer P to mPA under the notably mild TEA/oxidant conditions.</li> <li>The key structure-property relationships in mPA that determine its electronic characteristics (LUMO level, n-type behavior) and potential for electrical conductivity (backbone planarity).</li> <li>The overall significance of this approach for developing n-type conjugated polymers.</li> </ol>
<p><b>Sample Physics Research Subtask:</b>  For many physical phenomena involving plasmas, the plasmas can be treated as one or more fluids obeying fluid mechanics and Maxwell's laws. However, some analysis of a plasma requires that the different species of ions and electrons be treated kinetically with an assigned distribution function in phase space.</p> <p>Additionally, it is common to imagine the distribution function of an element of the plasma, say the electrons, as being anisotropic. Because a magnetized plasma has, by its very nature, a preferred direction (the direction of the magnetic field), it is entirely possible that there be an anisotropy between those species traveling parallel and those species traveling perpendicular to the magnetic field. One of the features that becomes visible when treating plasmas kinetically is the emergence of Bernstein modes. These are wave modes that travel at frequencies that are higher resonances of the plasma frequency of a species.</p> <p>Consider a plasma where the electrons are anisotropic such that <math>T_{\parallel} \neq T_{\perp}</math>. Consider electrostatic wave modes traveling perpendicular to the magnetic field with the condition on the upper hybrid frequency that <math>\Omega_e &lt; \omega_{\text{UH}} \leq 2\Omega_e</math>.</p> <p>First, determine the dispersion relation for an arbitrary <math>k_{\perp}</math>. You need only consider the first two harmonics. With this result, recover the cold plasma dispersion relation and the first term in the thermal correction.</p> <p>Next, consider a small parallel component to the wave number (<math>k_{\parallel}</math>), and find the condition for instability.</p>
<p><b>Sample Biology Research Subtask:</b>  The engineering of multicellular organisms to exhibit programmable behaviors, such as sophisticated environmental sensing, computation, and tailored responses, represents a grand challenge in synthetic biology. Achieving this requires the stable integration of complex genetic payloads into host genomes and their subsequent dynamic, multi-input regulation. Fundamental biological principles underpinning gene expression, DNA replication and repair, intercellular communication, and cellular resource management must be meticulously considered to design robust and predictable synthetic systems. These systems often need to operate orthogonally to, yet sometimes interface with, the host's endogenous regulatory networks across various developmental stages and physiological conditions.</p> <p>A research team aims to engineer a model plant (e.g., <i>Arabidopsis thaliana</i>) to produce a novel, high-value, three-enzyme metabolic pathway (Enzyme A, Enzyme B, Enzyme C, which must act sequentially) leading to metabolite X. The production of metabolite X needs to be tightly controlled, only occurring when both an abiotic stress signal (e.g., elevated salinity, sensed by an endogenous stress-responsive promoter <math>P_{\text{stress}}</math>) and the presence of a specific developmental cue (e.g., flowering, sensed by an endogenous flower-specific promoter <math>P_{\text{flower}}</math>) are detected. Furthermore, once metabolite X is produced in a specific cell, this cell must signal to its immediate, non-transgenic neighbors to upregulate a generic defense gene (<math>G_{\text{defense}}</math>) as a localized protective measure.</p> <p>Considering the fundamental principles of eukaryotic gene expression, genome stability, intercellular signaling, and metabolic engineering, propose and critically evaluate a comprehensive design strategy for such a system.</p>

**Figure 2: Sample FrontierScience-Research problems.** For the Research set, all experts hold a relevant PhD degree. The corresponding rubrics to these sample tasks can be found in Appendix A.

reasoning questions, particularly at the level of self-contained olympiad problems, but are still far from saturation on research-style work.

## 2 BENCHMARK CONSTRUCTION

### 2.1 DATA COLLECTION PIPELINE

FrontierScience-Olympiad questions were created in collaboration with 42 former international medalists or national team coaches in physics, chemistry, or biology, who have achieved 108 olympiad medals in total (45 gold, 37 silver, 26 bronze). All medalists were awarded from at least one (and often multiple) of the following olympiads: the International Physics Olympiad, International Chemistry Olympiad, International Olympiad on Astronomy and Astrophysics, European Physics Olympiad, and International Mendeleev Chemistry Olympiad.

FrontierScience-Research questions were created in collaboration with 45 qualified scientists and domain experts. The scientists were either post-doctoral researchers, professors, or doctoral candidates, often from globally recognized institutions. Qualitatively, each task was designed to represent a subproblem a PhD researcher might need to solve during the course of their research, and take at least three to five hours to successfully complete.

The scientists' areas of expertise spanned an array of scientific disciplines, including but not limited to: quantum mechanics, astrophysics, theoretical and experimental physics, biophysics, nanotechnology; molecular, evolutionary, and developmental biology, pharmacology, genomics, immunology, and neuroscience in biology; and biochemistry, physical and organic chemistry, materials and computational chemistry, catalysis, and photochemistry in chemistry. Experts are actively engaged in research for their domain with deep familiarity of research methodologies.

Each scientist wrote original problems for their track adhering to the following guidelines:

Task development pipeline

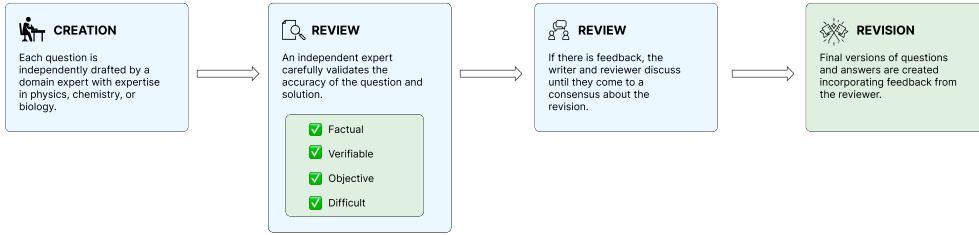


Figure 3: Tasks go through four stages: Creation, Review, Resolution, Revision. Independent experts review each other tasks to verify it aligns with the criteria.

	Olympiad	Research
<b>Originality</b>	<ul style="list-style-type: none"> <li>Problems are designed to mimic olympiad style challenges of complex, closed-form reasoning tasks.</li> <li>To minimize contamination risks, all problems are novel. While problems could draw initial inspiration from existing scientific ideas or questions, problems need to be re-contextualized for measuring reasoning capabilities through creative and non-obvious combinations or modifications.</li> </ul>	<ul style="list-style-type: none"> <li>Problems are designed to represent authentic scientific research tasks grounded in active areas of inquiry by our contributors. The specific problems will not be published outside of this dataset.</li> <li>To minimize contamination risks, all problems are expected to be novel. For problems that draw on the same core phenomena as published areas of study, or inspired by other problems, originality is enforced through the review process to ensure no direct overlap.</li> </ul>
<b>Difficulty</b>	<ul style="list-style-type: none"> <li>Problems at least the level of difficulty of international olympiad questions.</li> <li>Preliminary questions were evaluated against various internal models, where if the model answered correctly, the question was considered invalid and required an update.</li> </ul>	<ul style="list-style-type: none"> <li>Preliminary questions were evaluated against various internal OpenAI models. If the model scored highly on the rubric, the question was either discarded or significantly modified.</li> <li>Problems were calibrated such that 7-8 points out of 10 on the rubric was considered a successful solution.</li> </ul>
<b>Verifiability</b>	<ul style="list-style-type: none"> <li>The question provides all necessary variables, units, and information used in the final answer.</li> <li>The answer should be a single numeric or algebraic expression (physics and chemistry) or a fuzzy string-matchable answer (biology).</li> </ul>	<ul style="list-style-type: none"> <li>Each question includes a scoring rubric with multiple independent and objectively assessable items, totaling 10 points.</li> <li>Model judgments did not meaningfully differ from human judgments when using the rubric.</li> </ul>

For each research and olympiad problem, scientists provided a detailed solution that would earn full credit, as well as associated metadata (subdomains, difficulty levels, and sources of inspiration). Each contributed problem then went under review by at least one peer domain expert (for Research, each problem went under at least two reviews), who evaluated all components of the question against the guidelines. Questions could be inspired by known problems, or reference past work, but the guidelines were to make the task still new.

---

## 2.2 VERIFICATION PIPELINE

All submitted questions underwent an iterative review process. Independent domain expert reviewers read through each question, answer (either short answer for Olympiad, or rubric for Research), and solution explanation. The reviewers verified that each question was correct and followed all of the guidelines. The task creation process included some selection against OpenAI internal models (e.g., discarding tasks that models successfully got right, so we expect the evaluation to be biased against these models relative to others).

If any disagreements arose between the question writer and reviewer, they either came to a consensus or the question was discarded. Only after both experts agreed was the question submitted and added to the dataset. Experts for each domain in set then did a final review over each question in the submitted dataset, ensuring that all questions aligned with the guidelines.

For the Olympiad set, all problems went through at least one independent review, and then a holistic review by experts. For the Research set, all problems went through at least two independent reviews, and then a meta review by the experts. We increased review coverage for Research due to the questions being open-ended and rubrics being a newer and more imprecise grading architecture.

From over 500 Olympiad questions and over 200 Research questions, we did a meta-review with experts to filter down to an open-sourced gold set of 100 Olympiad questions and 60 Research questions. We keep the rest of the questions held-out to track potential contamination of the open-sourced set.

## 2.3 RUBRIC-BASED GRADING

The Olympiad set is gradable with a number, expression, or fuzzy string match, which improves verification. However, this verification often trades off with the expressivity and open-endedness of the problem. For the Research set, we introduce an experimental rubric-based architecture for grading more open-ended tasks.

Each question includes a scoring rubric with multiple independent and objectively assessable items, totaling 10 points. Each rubric item contains a description for a specific pass/fail condition (e.g., “Writes the following equation X”) and points. The grading rubric assesses not only the accuracy of the final answer, but also the correctness of intermediate reasoning steps, allowing for nuanced model performance and failure analysis. Scoring seven out of 10 points is considered a suitable solution and marked as a success. Due to the experimental design, we expect the Research set to have a lower noise ceiling than the Olympiad set. The flexibility of rubric points also enables other future grading procedures, such as average rubric points or different thresholds for what is considered a “success”.

Each question is accompanied by an explanatory solution path crafted by subject-matter experts. To run these evaluations without requiring human expert graders, we rely on judge models that assign a score given an attempted answer and a rubric. We provide model judge prompts in Appendix B that we use for all evaluations in this paper. We use GPT-5 at high reasoning effort for the model judge.

## 2.4 BENCHMARK COMPOSITION

FrontierScience contains a diverse range of scientific questions (Fig. 5). The Olympiad set is grounded in topics common on international science olympiad exams and is more weighted toward physics and chemistry over biology because it’s more feasible to develop questions that resolve to verifiable expressions and numbers. The Research set is grounded in contributors’ research specialties, with the gold set of 60 questions equally split between physics, chemistry, and biology.

## 3 EXPERIMENTS

### 3.1 MAIN RESULTS

We evaluated several frontier models: GPT-4o, OpenAI o4-mini, OpenAI o3, GPT-5.2, Claude Opus 4.5, Gemini 3 Pro, Grok 4, GPT-5.1, and GPT-5 on FrontierScience-Olympiad and FrontierScience-Research. All reasoning models were evaluated at “high” reasoning effort with the exception of GPT-

Sample Chemistry Research Problem	Sample Grading Rubric
<p>Phthalocyanines are important macrocyclic compounds widely utilized in material and life sciences due to their outstanding physicochemical properties, including intense absorption in the visible–near-IR region and high electron-accepting ability. However, unlike porphyrins, the structural diversity of phthalocyanines has been limited by traditional synthetic methods that rely on one-step condensation of isopindoline units. The development of methodologies that allow controlled modification of the phthalocyanine framework, particularly at the meso-positions, represents a significant advancement in expanding the structural and electronic diversity of these compounds.</p> <p>Analyze how the modification of meso-nitrogen atoms in nickel(II) phthalocyanine affects the electronic structure, aromaticity, spectroscopic properties, and reactivity of the resulting macrocycles. Explain the synthetic methodology that enables these modifications, the relationship between <math>\pi</math>-electron count and observed properties, and how these systems compare to traditional phthalocyanines and related porphyrinoids.</p>	<p><b>Analysis of Traditional Phthalocyanine Synthesis Limitations (1 point)</b></p> <p><input checked="" type="checkbox"/> 1.0 point: Correctly explains that traditional phthalocyanine synthesis involves one-step condensation with simultaneous formation of all meso-nitrogen bridges, providing limited control over substitution position at the periphery.</p> <p>0.5 point: Mentions limitations of traditional methods but without specific focus on meso-position control challenges.</p> <p>0.0 point: Fails to identify key limitations of traditional synthetic approaches or provides incorrect analysis.</p>
Sample Response	<p><b>Thiolate-Mediated Tetramerization Process (1 point)</b></p> <p>1.0 point: Correctly describes the thiolate-mediated reductive tetramerization and explains how counter cation size (<math>\text{K}^+</math> or <math>\text{Cs}^+</math> vs. <math>\text{Na}^+</math>) affects selectivity between tetramer formation and direct macrocyclization.</p> <p>0.5 point: Mentions thiolate-mediated tetramerization but without explaining factors controlling selectivity.</p> <p><input checked="" type="checkbox"/> 0.0 point: Incorrectly describes the oligomerization process or omits critical details about selectivity control.</p>
<p>1) Introduction: Why Modify the meso-Nitrogens of Phthalocyanines?</p> <p>Phthalocyanines (Pc) are a class of macrocyclic dyes known for their remarkable chemical, electronic, and optical properties. Their widespread use in materials science, catalysis, and biology stems from strong absorptions in the visible/near-IR regions, exceptional redox characteristics (i.e., high electron affinity), and rigid <math>\pi</math>-conjugation. Traditional synthetic methods, however, typically form the macrocycle by starting from phthalonitrile or substituted precursors, restricting the ability to introduce substitutions selectively at key ring positions – especially the “meso” (bridging) nitrogens. Recently, methodologies allowing targeted modification of the meso-nitrogen atoms in Ni(II) phthalocyanines have emerged as a powerful means to expand the structural and electronic diversity of these macrocycles.</p> <p>(... shortened for the purposes of this figure)</p> <p>2) Effect on Electronic Structure and Aromaticity</p> <p>a) <math>\pi</math>-Electron Count and Ring Currents</p> <p>Phthalocyanines are classical 18-<math>\pi</math>-electron aromatic systems, conferring high stability and planarity. Altering the meso-nitrogens through substitution or ring expansion can shift the effective <math>\pi</math>-electron count. For example, introducing additional unsaturated moieties or heteroatoms may expand the cyclic conjugation to 20 or 22 <math>\pi</math>-electrons. Changes to the electron count can shift the molecule from Hückel aromaticity (<math>4n + 2</math> <math>\pi</math>-electrons) to another aromatic or even antiaromatic regime, affecting ring currents and other properties (NMR chemical shifts, ring currents, etc.).</p> <p>b) Charge Distribution and Frontier Orbitals</p> <p>Meso-substitution can tune the electron density distribution around the macrocycle. Electron-donating groups can push electron density toward the ring center, lowering the energy of the lowest unoccupied molecular orbital (LUMO). Electron-withdrawing groups, conversely, stabilize the highest occupied molecular orbital (HOMO) and shift redox potentials toward more positive values. These changes alter both the electrochemical profile (more/less easily reduced or oxidized) and the Q and B-band positions in the UV-Vis spectrum.</p> <p>3) Spectroscopic Consequences</p> <p>a) UV-Vis Absorption (Q and B Bands)</p> <p>The principal absorption features of phthalocyanines lie in the visible (Q-band, typically 600–700 nm) and near-UV (B-band, typically 300–400 nm). Substitution that expands the ring conjugation or introduces electron-donating/withdrawing groups can:</p> <ul style="list-style-type: none"> <li>Shift the Q-band to longer wavelength (bathochromic shift) reaching into the near-IR, which is highly desirable for optoelectronic and photodynamic applications.</li> <li>Alter relative intensities of these bands and merge or split them, reflecting changes in orbital symmetries and energies.</li> </ul> <p>b) NMR Spectroscopy and Aromatic Ring Currents</p> <p>Modifications to the <math>\pi</math>-electron count and distribution are directly observed in <math>^1\text{H}</math> NMR (and <math>^{13}\text{C}</math> NMR) chemical shifts. More highly conjugated (or expanded) aromatic rings exhibit distinct downfield shifts for protons located within induced ring currents, while any partial loss of aromaticity or incorporation of antiaromatic segments can cause atypical shielding/deshielding patterns.</p> <p>4) Reactivity and Coordination Chemistry</p> <p>Because phthalocyanines are often used as redox catalysts or sensors, the meso-nitrogen modifications can significantly influence reactivity:</p> <ul style="list-style-type: none"> <li>Electron-rich meso substituents facilitate nucleophilic or electrophilic attacks at the ring periphery, enabling site-selective functionalizations that are otherwise difficult.</li> </ul> <p>(... shortened for the purposes of this figure)</p>	<p><b>Analysis of NMR Spectroscopic Features (1 point)</b></p> <p>1.0 point: Correctly explains that upfield shifts in the 16t system indicate para-<math>\pi</math>-ring current (antiaromaticity), contrasts this with the broad signals in 17t systems due to paramagnetism, and connects these observations to the underlying electronic structures.</p> <p><input checked="" type="checkbox"/> 0.5 point: Identifies basic NMR patterns but without clear connection to ring currents or electronic structure.</p> <p>0.0 point: Incorrectly interprets NMR data or fails to connect spectral features to electronic properties.</p>
	<p><b>Electrochemical Property Analysis (1 point)</b></p> <p>1.0 point: Correctly explains that the 16t system shows two reversible reductions reflecting conversion to 17t radical and 18t aromatic states, while 17t systems show narrow redox gaps due to facile interconversion between 16t, 17t, and 18t states, and relates these patterns to the underlying electronic structures.</p> <p><input checked="" type="checkbox"/> 0.5 point: Describes redox patterns without clearly connecting them to specific electronic state changes.</p> <p>0.0 point: Incorrectly interprets electrochemical data or fails to connect redox behavior to electronic properties.</p>
	<p><b>Analysis of Absorption Spectroscopy (1 point)</b></p> <p>1.0 point: Correctly explains that the 16t system shows weak/broad absorption due to symmetry-forbidden HOMO-LUMO transitions in antiaromatic systems, while 17t systems show Q-like bands plus NIR-II absorptions characteristic of radical species, and contrasts these with typical phthalocyanine spectral features.</p> <p><input checked="" type="checkbox"/> 0.5 point: Describes absorption features but provides limited connection to underlying electronic structures.</p> <p>0.0 point: Incorrectly interprets absorption data or fails to relate spectral features to electronic properties.</p>
	<p><b>Reactivity Analysis of Antiaromatic System (1 point)</b></p> <p>1.0 point: Correctly explains the high reactivity of the 16t system toward nucleophiles, details specific reactions with hydroxide (ring opening) and hydrazine (ring expansion), and explains how these transformations relieve strain in the system.</p> <p>0.5 point: Mentions reactivity but provides limited analysis of specific transformations or the driving forces behind them.</p> <p><input checked="" type="checkbox"/> 0.0 point: Incorrectly analyzes reactivity patterns or fails to connect them to the antiaromatic character of the 16t system.</p>
(... and more)	

Figure 4: Each task in the research set is graded using a rubric totaling 10 points that can be used by an expert or a judge model. To scale our ability to evaluate models, we use another model to judge responses.

5.2 at “xhigh”, and without browsing. In our initial evaluations, GPT-5.2 is the top performing model on FrontierScience, scoring 77% on the Olympiad Set and 25% on the Research set. Surprisingly, GPT-5 outperforms GPT-5.1 on the Research set and ties GPT-5.2. Overall, we’ve seen substantial progress on solving expert-level questions while leaving headroom for more progress, especially on open-ended research-style tasks.

For both sets, we use a model-based judge of GPT-5 (with a reasoning effort of “high”) to evaluate the models. For Olympiad, we give the judge the attempted answer and the actual answer and ask it to compare equivalency of the expression, number, or phrase. For Research, we give the judge the attempted answer and the rubric and ask it to return a single number reflecting the number of rubric points the answer earned (Appendix B).

When separating by subject, models perform comparably across distributions. For the Olympiad set, models perform better on chemistry, followed by physics and biology. For the Research set, models perform better on chemistry, followed by biology and then physics. Analyzing the transcripts, models typically struggled with reasoning or logic error, failures in understanding niche concepts, calculation errors, and factual inaccuracy.

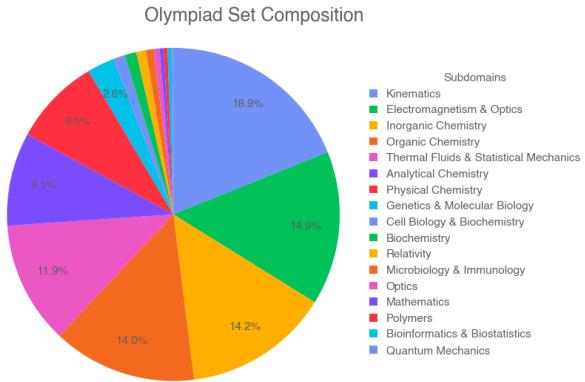


Figure 5: The Olympiad split is composed of a diverse set of topics, from biochemistry to quantum mechanics.

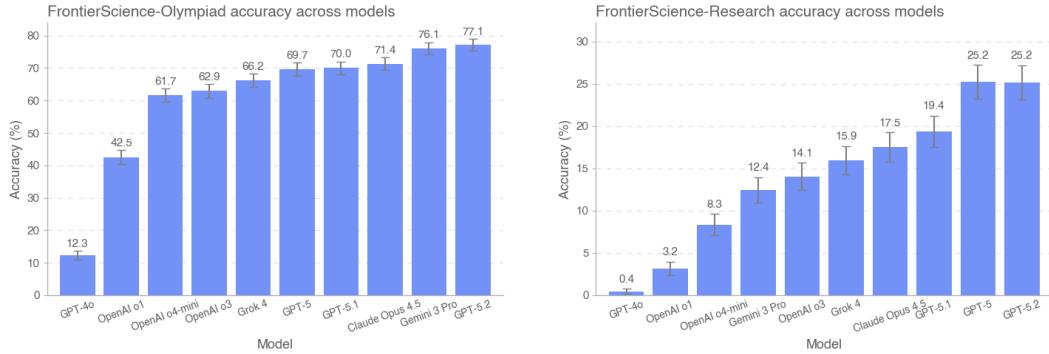


Figure 6: We compare accuracies across several frontier models. GPT-5.2 is our highest performing model across the Olympiad and the Research set. Gemini 3 Pro is comparable to GPT-5.2 on Olympiad, and GPT-5 is tied with GPT-5.2 on Research. For all Olympiad evaluations, scores were averaged across 20 independent trials. For all Research evaluations, scores were averaged across 30 independent trials, using a threshold of a response earning at least seven rubric points as correct.

## 4 DISCUSSION

While FrontierScience represents an advance in understanding scientific research capabilities, there are multiple limitations:

1. **Constrained problem-solving:** A significant part of scientific research is proposing novel research directions, hypotheses, and ideas. FrontierScience is composed of questions with a constrained problem statement, which focuses on evaluating the reasoning to complete a research task and less on ideation. While the Research set aims to measure more open-ended reasoning, this is an inherent limitation of an autogradable Q&A style evaluation.
2. **Rubric reliability:** We sought to improve rubric reliability of the Research set through strict guidelines, verification, and consistency with human grading. However, the rubric is less objective than the equivalency checker of a single expression or number, and relies on the model judge's capabilities.
3. **Modalities:** Problems are designed to be text-only without image or video outputs. Modalities beyond text are more representative of scientific research. In particular, real-world scientific research often involves interaction with reality (e.g., wet labs), which this evaluation does not cover.

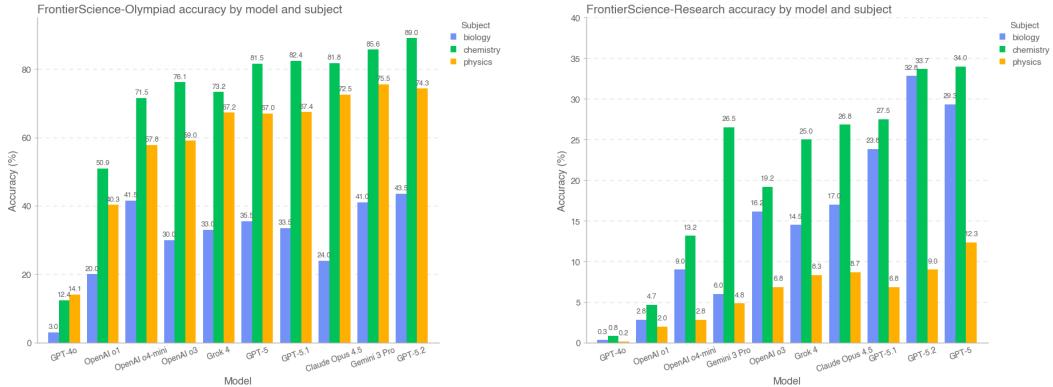


Figure 7: We compare accuracies across several frontier models on FrontierScience-Olympiad (Left) and FrontierScience-Research (Right) with accuracies split out by subject.

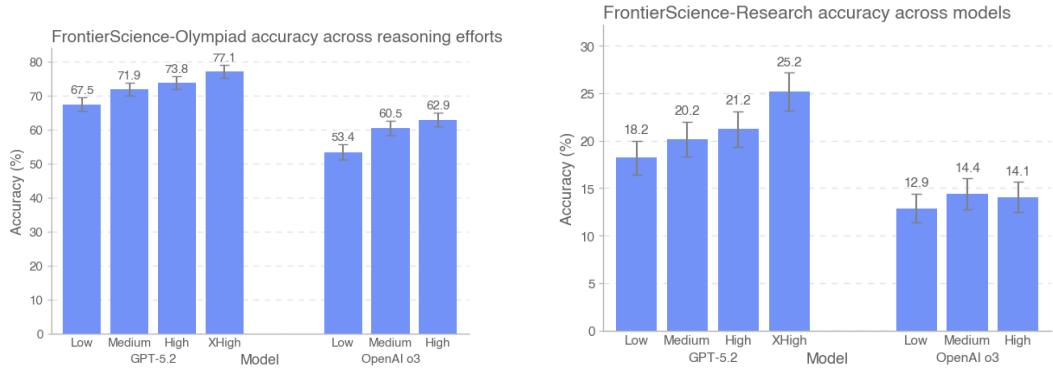


Figure 8: We compare accuracies for GPT-5.2 and OpenAI o3 on FrontierScience-Olympiad and FrontierScience-Research across different reasoning efforts. Using more test-time tokens enables GPT-5.2 to go from 67.5% to 77.1% on the Olympiad set, as well as from 18% to 25% on the Research set. Surprisingly, o3 performs marginally worse at high reasoning effort compared to medium reasoning effort on the Research set.

4. **Human baselining:** We did not perform human baselining of this dataset and leave that to future work. Since the questions are grounded in experts’ authentic research, an interesting question is how to conduct a human baseline. Since the questions are so specialized, it may be important to find experts in that speciality to solve them and derive a consensus baseline.

Research and practical evaluations will be important to continue building long standing and directly relevant evaluations. Scientific reasoning is important for the beneficial impacts of AI and we hope for continued development of robust and relevant benchmarks for accelerating scientific progress.

## 5 RELATED WORK

Previous science-oriented and knowledge benchmarks primarily measure models’ capabilities through multiple-choice or single-answer formats. Benchmarks such as MMLU (Hendrycks et al., 2021; Wang et al., 2024b), GPQA (Rein et al., 2023), and ScienceQA (Lu et al., 2022) have significantly contributed to understanding model performance across scientific knowledge and basic reasoning tasks. However, these benchmarks largely target knowledge retrieval or recognition of well-known scientific concepts rather than research-level scientific reasoning. GPQA, for instance, measures general-purpose science reasoning but is limited to structured multiple-choice settings, reducing diagnostic power of more complex, open-ended tasks.

---

To evaluate more advanced reasoning skills, recent benchmarks have introduced open-ended questions. OlympiadBench (He et al., 2024) introduced high-school-level Science Olympiad-style questions, demonstrating the value of open-ended and auto-gradable formats. However, it is focused on collecting pre-existing math and physics questions, raising contamination concerns. The FrontierScience Olympiad track extends this format by employing international Olympiad medalists across a range of scientific subjects to craft problems specifically adversarial against state-of-the-art models. Complementary work such as PHYBench (Qiu et al., 2025), ChemBench Mirza et al. (2024), and SciBench (Wang et al., 2024a) also extend constrained reasoning tasks to certain domains. Other benchmarks, such as PaperBench (Starace et al., 2025) investigate capabilities on AI research tasks for replicating papers, while being less focused on scientific capabilities.

CritPt (Zhu et al., 2025) introduces a PhD-level physics benchmarks that focuses on difficult, unpublished research questions, employing a methodology of verifiable checkpoints. FrontierScience trades off on the benefits of fully verifiable checkpoints to evaluate more open-ended research subtasks, which is also shown by its extension to chemistry and biology questions. LAB-Bench (Laurent et al., 2024) is a broad and diverse benchmark of biology questions that are relevant to practical workflows. It focuses on multiple choice questions across skills such as recalling literature and manipulating DNA and protein sequences. FrontierScience is complementary to this work by focusing on difficult reasoning questions rather than day-to-day scientific workflows.

Previous approaches to open-ended problem evaluation typically rely on final-answer correctness as a primary assessment metric, limiting the insight into intermediate reasoning steps. Prior work such as the LLM-Rubric (Hashemi et al., 2024) have incorporated rubric-based evaluations for evaluating LLM responses in dimensions of naturalness, conciseness, etc, and recent work such as HealthBench (Arora et al., 2025) has used this format in real-world relevant domains. The FrontierScience Research benchmark builds upon the structured rubric-based evaluation approach to test for reasoning with custom rubric items per evaluation. Each rubric item in FrontierScience Research problems explicitly decomposes answers into granular components, enabling more nuanced analyses of where and why models succeed or fail, particularly valuable given the complexity of PhD-level scientific research tasks.

## 6 ACKNOWLEDGMENTS

We sincerely thank our external partners and expert evaluators for their valuable contributions, including their time, domain expertise, and thoughtful feedback.

We thank Addea Gupta, Alex Karpenko, Andy Applebaum, Bowen Jiang, David Robinson, Elizabeth Proehl, Evan Mays, Grace Kim, Ilge Akkaya, Jerry Tworek, Joy Jiao, Kevin Liu, Leon Maksin, Leyton Ho, Michele Wang, Michele Wang, Nat McAleese, Nikolai Eroshenko, Olivia Watkins, Patrick Chao, Phillip Guo, Phoebe Thacker, Rahul Arora, Ryan Kaufman, Samuel Miserendino, Sebastian Bubeck, Simón Fishman, Stephen McAleer, and Ven Chandrasekaran for helpful discussions, feedback, and support.

## REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025. URL <https://arxiv.org/abs/2505.08775>.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.745. URL <http://dx.doi.org/10.18653/v1/2024.acl-long.745>.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

---

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*, 2021.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnappati, Andrew D. White, and Samuel G. Rodrigues. Lab-bench: Measuring capabilities of language models for biology research, 2024. URL <https://arxiv.org/abs/2407.10362>.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, and Kai-Wei Chang. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martíño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. Are large language models superhuman chemists?, 2024. URL <https://arxiv.org/abs/2404.01475>.

OpenAI. Introducing gpt-5.2. <https://openai.com/index/introducing-gpt-5-2/>, 2025. Accessed: 2025-12-15.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhan Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. Phybench: Holistic evaluation of physical perception and reasoning in large language models, 2025. URL <https://arxiv.org/abs/2504.16074>.

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. *arXiv preprint arXiv:2311.12022*, 2023.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating ai's ability to replicate ai research, 2025. URL <https://arxiv.org/abs/2504.01848>.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models, 2024a. URL <https://arxiv.org/abs/2307.10635>.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024b. URL <https://arxiv.org/abs/2406.01574>.

Minhui Zhu, Minyang Tian, Xiaocheng Yang, Tianci Zhou, Lifan Yuan, Penghao Zhu, Eli Chertkov, Shengyan Liu, Yufeng Du, Ziming Ji, Indranil Das, Junyi Cao, Yufeng Du, Jiabin Yu, Peixue Wu, Jinchen He, Yifan Su, Yikun Jiang, Yujie Zhang, Chang Liu, Ze-Min Huang, Weizhen Jia, Yunkai Wang, Farshid Jafarpour, Yong Zhao, Xinan Chen, Jessie Shelton, Aaron W. Young, John Bartolotta, Wenchao Xu, Yue Sun, Anjun Chu, Victor Colussi, Chris Akers, Nathan Brooks, Wenbo Fu, Jinchao Zhao, Marvin Qi, Anqi Mu, Yubo Yang, Allen Zang, Yang Lyu, Peizhi Mai, Christopher Wilson, Xuefei Guo, Juntai Zhou, Daniel Inafuku, Chi Xue, Luyu Gao, Ze Yang, Yaïr

---

Hein, Yonatan Kahn, Kevin Zhou, Di Luo, John Drew Wilson, Jarrod T. Reilly, Dmytro Bandak, Ofir Press, Liang Yang, Xueying Wang, Hao Tong, Nicolas Chia, Eliu Huerta, and Hao Peng. Probing the critical point (critpt) of ai reasoning: a frontier physics research benchmark, 2025. URL <https://arxiv.org/abs/2509.26574>.

## APPENDIX

### A FULL SAMPLE RESEARCH PROBLEMS

**Sample Physics Research Problem:**

For many physical phenomena involving plasmas, the plasma can be treated as one or more fluids obeying fluid mechanics and Maxwell's laws. However, some analysis of a plasma requires that the different species of ions and electrons be treated kinetically with an assigned distribution function in phase space.

Additionally, it is common to imagine the distribution function of an element of the plasma, say the electrons, as being anisotropic. Because a magnetized plasma has, by its very nature, a preferred direction (the direction of the magnetic field), it is entirely possible that there is anisotropy between those species traveling parallel and those species traveling perpendicular to the magnetic field.

One of the features that becomes visible when treating plasmas kinetically is the emergence of Bernstein modes. These are wave modes that travel at frequencies that are higher resonances of the plasma frequency of a species.

Consider a plasma where the electrons are anisotropic such that  $T_{\parallel} \neq T_{\perp}$ . Consider electrostatic wave modes traveling perpendicular to the magnetic field with the condition on the upper hybrid frequency that  $\Omega_{\text{h}} < \omega_{\text{UH}} \leq 2\Omega_{\text{e}}$ .

First, determine the dispersion relation for an arbitrary  $k_{\perp}$ . You need only consider the first two harmonics. With this result, recover the cold plasma dispersion relation and the first term in the thermal correction.

Next, consider a small parallel component to the wave number ( $k_{\parallel}$ ), and find the condition for instability.

**Sample Physics Research Rubric**

**Points: 1.0, Item:** Analysis restricted to the first two harmonics (only two terms in the infinite sum are taken).

**Points: 1.0, Item:** Correctly find the second harmonic wave frequency with the first thermal correction,

$$(\omega/\Omega_{\text{h}})^2 = 4 + 3(\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)/(3\Omega_{\text{h}}^2 - \omega_{\text{pe}}^2) \cdot \mu_{\text{pe}}$$

**Points: 1.0, Item:** Correctly finds the dispersion relation for a general temperature,

$$(\omega/\Omega_{\text{h}})^2 = (1/2)[5 + 2(\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)(\Gamma_{\perp}(\mu_{\text{pe}}) + 4\Gamma_{\parallel}(\mu_{\text{pe}}))]/\mu_{\text{pe}} + \sqrt{9 - 12(\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)(\Gamma_{\perp}(\mu_{\text{pe}}) + 4\Gamma_{\parallel}(\mu_{\text{pe}}))}/(\Gamma_{\perp}(\mu_{\text{pe}}) + 4\Gamma_{\parallel}(\mu_{\text{pe}}))/\mu_{\text{pe}}^2].$$

**Points: 1.0, Item:** Correctly finds the first harmonic wave frequency with its thermal correction,

$$(\omega/\Omega_{\text{h}})^2 = 1 + (\omega_{\text{pe}}^2/\Omega_{\text{h}}^2) - 3(\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)/(3\Omega_{\text{h}}^2 - \omega_{\text{pe}}^2) \cdot \mu_{\text{pe}}$$

**Points: 1.0, Item:** Correctly gives the general form for the imaginary part of the dielectric constant,

$$\text{Im}[\epsilon] = (\sqrt{n}/(k^2\lambda)^2) \cdot \sum_{\perp} (\Gamma_{\perp}(\mu_{\text{pe}}) \cdot (\omega + A\Omega_{\text{h}})/(k\|\mathbf{a}\|_{\perp}) - \epsilon^2 - ((\omega - i\Omega_{\text{h}})/(k\|\mathbf{a}\|_{\perp}))^2).$$

**Points: 1.0, Item:** Correctly simplifies the dispersion relation for the cold plasma limit,

$$(\omega/\Omega_{\text{h}})^2 = (1/2)[5 + (\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)] \pm \sqrt{9 - 6(\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)(1 - 2\mu_{\text{pe}}) + 4(\omega_{\text{pe}}^2/\Omega_{\text{h}}^2)}.$$

**Points: 1.0, Item:** Correctly simplifies the imaginary part of the dielectric for the first two harmonics,

$$\text{Im}[\epsilon] = (\sqrt{n}/(k^2\lambda)^2) \cdot (X_{\perp}(\omega + A\Omega_{\text{h}})/(k\|\mathbf{a}\|_{\perp}) + X_{\parallel}(\omega + 2A\Omega_{\text{h}})/(k\|\mathbf{a}\|_{\perp})).$$

**Points: 1.5, Item:** Derives the correct condition for wave instability,

$$\Lambda < -(\omega/\Omega_{\text{h}}) \cdot (X_{\perp}X_{\parallel}/(X_{\perp} + 2X_{\parallel})) < 0.$$

**Points: 1.0, Item:** Includes the correct equation for the dispersion relation for this situation,

$$\epsilon = 1 - (1/(k^2\lambda^2)) \cdot \sum_{\perp} (\Gamma_{\perp}(\mu_{\text{pe}}) \cdot [2(\omega^2/\Omega_{\text{h}}^2)/(\omega^2 - k^2\Omega_{\text{h}}^2)] - r_{\perp}(\mu_{\text{pe}}).$$

**Points: 0.5, Item:** Recognizes that a negative imaginary part in the susceptibility will drive wave instability.

Figure 9: Sample FrontierScience-Research Physics Problem.

**Sample Biology Research Problem:**

The engineering of multicellular systems to exhibit programmable behaviors, such as sophisticated environmental sensing, computation, and tailored responses, represents a grand challenge in synthetic biology. Achieving this requires the safe integration of complex genetic payloads into host genomes and their subsequent dynamic, multi-input regulation. Fundamental biological principles underpinning gene expression, DNA replication and repair, intercellular communication, and cellular resources management must be methodically considered to design robust and predictable synthetic systems. These systems often need to operate orthogonally to, yet sometimes interface with, the host's endogenous regulatory across various developmental stages and physiological conditions.

A research team aims to engineer a plant model (e.g., *Arabidopsis thaliana*) to produce a novel, high-value, three-enzyme metabolic pathway (Enzyme A, Enzyme B, Enzyme C, which must act sequentially) leading to metabolite X. The production of metabolite X needs to be tightly controlled, only occurring when both an abiotic stress signal (e.g., elevated salinity, sensed by an endogenous stress-responsive promoter  $P_{\text{stress}}$ ) and the presence of a specific developmental cue (e.g., flowering, sensed by an endogenous flower-specific promoter  $P_{\text{flower}}$ ) are detected. Furthermore, once metabolite X is produced in a specific cell, this cell must signal to its immediate, non-transgenic neighbors to up-regulate a generic defense gene ( $G_{\text{defense}}$ ) as a localized protective measure.

Considering the fundamental principles of eukaryotic gene expression, genome stability, intercellular signaling, and metabolic engineering, propose and critically evaluate a comprehensive design strategy for such a system.

**Sample Biology Research Problem**

**Points: 1.0, Item:** Details the metabolic pathway expression cassette architecture as a polycistronic-like unit utilizing self-cleaving 2A peptides for co-expression of Enzymes A, B, and C.

**Points: 1.0, Item:** Justifies the transposase-assisted, CRISPR-Cas9 integration by citing its potential for "seamless" integration with defined target site duplications (e.g., TTAAT) and reduced likelihood of deletions/rearrangements compared to simple NHEJ knock-in.

**Points: 1.0, Item:** Proposes a CRISPR-based split-activator system (e.g., dCas9 fused to N- and C-terminal halves of an activator like VP64) for the AND logic gate.

**Points: 1.0, Item:** Proposes engineering the mobile peptide signal for enhanced stability (e.g., substituting protease-labile amino acids, cyclization, or fusion to a stable carrier) to combat degradation/sequestration. This mobile peptide must get secreted into the apoplast to diffuse to neighboring cells and signal for defense upregulation.

**Points: 1.0, Item:** Proposes flanking the entire synthetic transposon with chromatin insulator elements to specifically address epigenetic instability of the integrated cassette

**Points: 1.0, Item:** Proposes incorporating additional layers of repression (e.g., a dCas9-repressor fusion also targeting  $P_{\text{output}}$ , with its own complex regulation) to tighten the AND gate's "OFF-state".

**Points: 1.0, Item:** Proposes protein stability engineering for Enzyme C or using inducible degradation tags for Enzymes A/B as specific strategies to address metabolic imbalance and intermediate cytotoxicity.

**Points: 1.0, Item:** Proposes sequence diversification of internal repetitive promoters and terminators within the synthetic construct to mitigate homology-dependent gene silencing, providing specific examples like using Nos or ubiquitin promoters instead of multiple CaMV 35S copies.

**Points: 1.0, Item:** Proposes the use of a synthetic, orthogonal peptide-receptor pair, engineered for high affinity and specificity, as the primary strategy to reduce off-target signaling.

**Points: 1.0, Item:** Provides a clear mechanistic explanation of the split-dCas9-activator AND gate, detailing how co-localization and reconstitution of the activator domain by two distinct input-driven sgRNAs (sgRNA1 from  $P_{\text{stress}}$ , sgRNA2 from  $P_{\text{flower}}$ ) leads to  $P_{\text{output}}$  activation.

Figure 10: Sample FrontierScience-Research Biology Problem.

### B EVALUATION PROMPTS

For our evaluations, we use a judge model based on GPT-5 thinking at high reasoning effort to judge responses. Here, we provide the exact prompts we give the judge model for the evaluations in this paper.

**Sample Chemistry Research Problem:**

The development of stable, high-conductivity n-type conjugated polymers is crucial for advancing organic electronics but lags behind p-type materials. Polyacetylene analogues are attractive targets, but incorporating electron-withdrawing groups to achieve low LUMO energies often disrupts backbone planarity essential for conductivity. Novel synthetic strategies are needed to create well-defined, planar, electron-deficient conjugated polymers.

Maleimide Polyacetylene (mPA), featuring an alternating vinylene (-CH=CH-) unit and N-alkylated maleimide unit backbone, is synthesized via a two-stage strategy:

1. ROMP: An N-alkylated maleimide-fused cyclobutene monomer (M) is polymerized using a Mo-based Schrock catalyst to yields a soluble, non-conjugated precursor polymer (P) containing alternating vinylene and N-alkylated dihydro-maleimide units.
2. Oxidation: The precursor P is converted to the fully conjugated mPA using triethylamine (TEA) and a mild oxidant (e.g., TCNQ or  $I_2$ ).

Provide a comprehensive chemical analysis of this system, addressing:

- The strategic rationale for employing the two-stage precursor ROMP approach and the specific catalyst choice.
- The complete mechanistic basis for the conversion of the precursor polymer P to mPA under the notably mild TEA/oxidant conditions.
- The key structure-property relationships in mPA that determine its electronic characteristics (LUMO level, n-type behavior) and potential for electrical conductivity (backbone planarity).
- The overall significance of this approach for developing n-type conjugated polymers.

**Sample Chemistry Research Rubric:**

**Points: 1.0, Item: Conductivity: Role of Planarity:** Explains the importance of backbone planarity for high conductivity by explicitly linking it to efficient  $\pi$ -orbital overlap enabling both effective intrachain charge delocalization AND favorable interchain charge transport (e.g., via  $\pi$ -stacking).

**Points: 1.0, Item: Electronic Structure:** Consequence of Low LUMO: Explains that the low LUMO energy level to both facilitating facile n-doping (reduction) (0.5 points) AND enhancing the electrochemical stability of the resulting negatively charged (anionic/polaronic) states on the polymer backbone (0.5 points).

**Points: 1.0, Item: Electronic Structure: LUMO Lowering Mechanism:** Attributes the lowering of the LUMO energy level in mPA primarily to the strong electron-withdrawing nature of the conjugated maleimide carbonyl groups (0.5 points), referencing both their inductive (-I) and resonance (-M) effects (0.5 points).

**Points: 1.0, Item: Mechanism: CT Complex Postulation:** Explicitly proposes the formation of a Charge-Transfer (CT) complex between TEA and the oxidant AND identifies this complex formation as the key element that allows the weakly basic TEA to effectively initiate the transformation despite the high pKa of the  $\pi$ -electrons.

**Points: 1.0, Item: Mechanism: Initial Activation Step:** Describes the initial step involving the CT complex interacting with polymer P to achieve activation of the C-H bond or transient/partial proton abstraction, emphasizing that full stoichiometric deprotection by TEA alone is unlikely and unnecessary in this proposed synergistic mechanism.

**Points: 1.0, Item: Mechanism: Redox Transformation & Oxidant Function:** Accurately identifies the  $P \rightarrow mPA$  conversion as a net two-electron, two-proton oxidation (dehydrogenation) per dihydro-maleimide repeat unit, involving the conversion of  $sp^2$  hybridized carbons alpha to the carbonyl to  $sp^3$  carbons to form the intramolecular C=C double bond within the maleimide ring.

Explicitly states that the thermodynamic driving force for this transformation is the formation of the highly stable, extended  $\pi$ -conjugated system along the polymer backbone.

Correctly identifies the role the oxidant (TCNQ or  $I_2$ ) as the stoichiometric terminal electron acceptor, specifying that it undergoes a two-electron reduction (per equivalent of double bond formed) to its corresponding reduced form (e.g., TCNQ<sup>-</sup> or  $I^-$ ), thereby enabling the overall redox transformation.

**Points: 1.0, Item: Molecular Structure: Planarity Analysis:** Correctly argues for the high degree of backbone planarity in mPA by referencing the inherent planarity of both the vinylene units and the maleimide rings, AND notes the absence of significant steric hindrance directly on the conjugated backbone that would force twisting.

**Points: 1.0, Item: Rationale for Catalyst Choice:** Correctly identifies the need for a high-activity ROMP catalyst for the strained cyclobutene monomer AND explicitly states that Mo-based Schrock catalysts possess significantly higher reactivity for such monomers compared to Ru-based Grubbs catalysts.

**Points: 1.0, Item: Rationale for Precursor Strategy:** Provides a comprehensive, chemically detailed rationale citing all three major advantages:

- Quantifies the processability advantage by explicitly linking the solubility of precursor P (due to its  $sp^3$  carbons disrupting conjugation/packing and the N-alkyl groups) to its suitability for solution-based processing techniques (e.g., film casting), contrasting this with the expected rigidity, strong interchain interactions, and poor solubility of the fully conjugated mPA target which would hinder direct processing (1/3 points)
- Explains that living/controlled ROMP provides precise control over  $M_n$  and low  $D$ , which is critical for optimizing charge transport in semiconductors (by ensuring chain uniformity and minimizing defects/end-groups) and device reproducibility, explicitly contrasting this with the typical lack of such control in Ziegler-Natta or other direct polyacetylene polymerizations (1/3 points)
- Highlights that the precursor route guarantees the regioregularity and specific alternating sequence defined by monomer M is perfectly preserved in mPA via selective backbone modification, avoiding potential side reactions or structural scrambling common in direct polymerization of functionalized acetylenes (1/3 points)

**Points: 1.0, Item: Significance:** Overcoming Specific n-Type Polymer Limitations:

Articulates the significance by explaining how the mPA system provides a synergistic solution to the often-competing demands (the "trilemma") in high-performance n-type polymer design. Must explicitly discuss how:

- the maleimide structure provides extreme electron deficiency (very low LUMO) necessary for air-stable n-doping and potentially favorable device energetics, addressing stability issues common in prior low-LUMO polymers
- the alternating vinylene-maleimide backbone architecture inherently maintains high planarity, overcoming the steric twisting often introduced by electron-withdrawing substituents in other PA analogues, thus facilitating efficient charge transport
- the ROMP precursor synthetic strategy enables architectural control ( $M_n$ ,  $D$ ) and processability, circumventing synthetic limitations faced by many rigid conjugated polymers.

Figure 11: Sample FrontierScience-Research Chemistry Problem.

### FrontierScience-Olympiad Judge Model Prompt

You are grading an attempted answer to a science olympiad problem. You will be given the problem, attempted answer, and reference answer. Evaluate the solution against the provided reference solution, ensuring it is complete and matches the reference solution. Pay close attention to detail and grade it strictly, but fairly.

The reference answer is either a single number or expression in latex formatting, a chemical formula, a compound name, or a phrase referring to a specific name, entity, or method.

Mark the attempted answer as correct if it fully matches the reference answer or is otherwise equivalent (e.g., an equivalent algebraic expression, a numerical number within 1 decimal place rounding of the reference answer (e.g.,  $6.69 \approx 6.7$ ), an equivalent name for a compound/formula, equivalent when accounting for units, etc.). Mark it as incorrect if it is not equivalent to the reference answer.

\*\*\*

The problem: {problem}

\*\*\*

The reference answer: {reference\_answer}

\*\*\*

The attempted answer: {answer}

\*\*\*

First, think step-by-step about whether the attempted answer matches the reference answer. If the attempted answer is correct, write "VERDICT: CORRECT" in the last line of your response, with no other text or formatting. If it is incorrect, write "VERDICT: INCORRECT".

---

### FrontierScience-Research Judge Model Prompt

You are grading a science exam.

You will be given the problem, attempted answer, and a rubric to grade the answer. The rubric will total up to 10 points.

Evaluate the attempted answer against the provided rubric. Pay close attention to detail and grade it strictly, but fairly. Only evaluate against the rubric, as you yourself should not make any judgements (e.g., even if you think the answer is correct but rubric is wrong, you should treat the rubric as the gold standard). Return the absolute total number of points earned (it can be a decimal based on the rubric). \*\*\*

The problem: {problem}

\*\*\*

The rubric: {rubric}

\*\*\*

The attempted answer: {answer}

\*\*\*

First, think step-by-step about each rubric item. Explain your reasoning for each rubric item. Then, tally the points up and write VERDICT: *{total points}* in the last line of your response, no other text. For example, VERDICT: 2.5 or VERDICT: 8.

## C PROBLEM REQUIREMENTS

We display a summarized list of requirements given to each problem writer for both the Olympiad and the Research set.

## Research Problem Guidelines

### Question clarity:

- All necessary background information, variables, notation and assumptions must be explicitly defined.
- The model should have access to the same inputs that would be necessary for an expert in the field to solve the question. If a model fails due to missing information, adjust the description or task wording.

### Originality:

- If the question is inspired by a publicly available source (e.g., research paper), the question and/or answer values should be substantially modified.

### Grading consistency:

- Rubric items must be independent and objectively assessable.
- Each rubric description should:
  - Be affirmative, clear, and explicitly state required conditions for credit, avoiding vagueness.
  - Provide specific pass/fail conditions (e.g., “Writes the following equation X”).
  - Define all variables and acronyms.
- Ensure discrepancies between human and model grading results do not exceed 0.5-points overall.

### Difficulty:

- Problems should be sufficiently challenging, typically requiring 3–5 hours to draft, to adequately test depth of reasoning.
- The questions should evaluate the model’s ability to reason at a complex level – researchers should test problem solving rather than prose, search or recency (knowledge cutoffs).

## Olympiad Problem Guidelines

### Correctness – ensure problem is correct and complete:

- Provide all assumptions and define all variables.
- Use correct physics/equations.
- Make sure the problems cannot be interpreted in multiple ways.

### Verifiable – ensure problem is verifiable:

- Provide all variables used in the final answer.
- Make sure information about directions/plus/minus signs is given.
- Clear LaTeX formulations – no Unicode characters and ensure all functions are formatted correctly (a common mistake is to write  $\sin$  rather than  $\sin$ ).
- Define a variable for what you ask for in the question (e.g., do not write “find an expression for the energy in terms of the following variables”; instead write “find an expression for  $E$ , the energy, in terms of the following variables”).
- Ensure the answer is a single numeric or algebraic expression.
- If using units in the final answer, explicitly specify the unit you want and what symbol to use (e.g., “give your final answer in units of meters using the symbol  $m$ ”).