SciRisk-Bench: A Risk-Dimension-Aware Benchmark for AI4Science Safety

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly embedded in AI for Science (AI4Science) workflows, from scientific question answering and literature analysis to laboratory planning and autonomous discovery. This progress creates an urgent need for safety benchmarks that evaluate not only scientific competence, but also whether models recognize and avoid risks in high-stakes scientific contexts. Existing AI4Science safety datasets cover several disciplines and task formats, leaving the underlying risk dimensions underspecified. We introduce **SciRisk-Bench**, a benchmark designed to evaluate AI4Science safety from two complementary perspectives: explicit risk dimensions and scientific disciplines. SciRisk-Bench covers 7 disciplines, 31 subdisciplines and 10 risk dimensions. In the experimental section, we evaluate both mainstream LLMs and science-oriented LLMs across risk dimensions, disciplines, and sub-disciplines, enabling fine-grained diagnosis of where scientific models remain unsafe.

1 Introduction

AI4Science has become a central paradigm for accelerating scientific discovery. Recent systems have demonstrated that machine learning and LLM-based methods can assist mathematical program search (Romera-Paredes et al., 2024) and discover efficient algorithms (Mankowitz et al., 2023). In materials science, AI has supported large-scale materials discovery (Merchant et al., 2023) and autonomous synthesis (Szymanski et al., 2023). In biology, protein structure prediction has been transformed by AlphaFold (Jumper et al., 2021), with later work extending biomolecular modeling to broader molecular complexes (Krishna et al., 2024). In geoscience, foundation models have been proposed for weather and climate modeling (Nguyen et al., 2023), and neural forecasting systems have achieved strong medium-range weather prediction (Lam et al., 2023). Foundation models are also entering generalist medical AI (Moor et al., 2023) and clinical knowledge reasoning (Singhal et al., 2023). As LLMs become natural-language interfaces to scientific knowledge, tools, and protocols, they increasingly mediate decisions that may affect laboratories, public health, critical infrastructure, and scientific governance.

This expanding role makes AI4Science safety a distinct and urgent evaluation problem. Scientific mistakes are not limited to ordinary factual errors: an unsafe answer may provide actionable dual-use details, omit laboratory precautions, overstate uncertain evidence, expose private or sensitive data, misrepresent regulations, or give authoritative-sounding but false explanations. Prior studies have shown that AI systems can amplify dual-use risks in drug discovery (Urbina et al., 2022) and rely on misleading shortcuts in medical imaging (DeGrave et al., 2021). Chemistry-specific prompting attacks further expose safety vulnerabilities in molecular representations (Wong et al., 2024), while synthetic biology and AI convergence raises broader regulatory and security concerns (Hynek, 2025). General LLM safety benchmarks are useful, but scientific settings require specialized evaluation because risk is tightly coupled with domain expertise, experimental context, and regulatory constraints.

Several benchmarks have begun to address this gap. SciBench evaluates college-level scientific problem solving (Wang et al., 2023), ScienceQA focuses on multimodal science question answering (Lu et al., 2022), SciEval targets multi-level scientific research evaluation (Sun et al., 2024), and SciKnowEval measures multi-level scientific knowledge (Feng et al., 2024). Safety-oriented efforts have also emerged: ChemSafetyBench targets chemistry safety (Zhao et al., 2024), MedSafetyBench evaluates harmful medical requests (Han et al., 2024), LabSafetyBench focuses on laboratory safety (Zhou et al., 2024), SciSafeEval evalu-

ates scientific safety alignment (Li et al., 2024b), WMDP measures malicious-use knowledge (Li et al., 2024a), SOSBench studies safety alignment on scientific knowledge (Jiang et al., 2025), and SafeScientist evaluates risk-aware scientific agents (Zhu et al., 2025). However, most existing benchmarks still emphasize either disciplinary coverage or broad safety categories. They provide limited visibility into which *types* of scientific risk drive unsafe behavior inside each discipline.

We propose **SciRisk-Bench**, a risk-dimension-aware benchmark for AI4Science safety. SciRisk-Bench spans seven scientific disciplines, including domains such as biology, chemistry, geography, engineering, and physics, with representative sub-disciplines ranging from synthetic biology and organic synthesis to GIS and nuclear physics. The full discipline hierarchy is described in the Method section. Unlike prior work that primarily treats scientific safety as a domain-level problem, SciRisk-Bench explicitly annotates examples by risk dimensions. For example, dual-use captures scientific knowledge that can enable harmful misuse, laboratory safety concerns missing precautions in experimental settings, and hallucinations and misconceptions cover confident but false scientific claims. This design enables evaluation to answer not only "which discipline is unsafe?", but also "which risk mechanism causes the failure?"

Our experiments evaluate mainstream LLMs and science-oriented LLMs across risk dimensions, disciplines, and sub-disciplines, showing that science-specialized models often exhibit higher ASR despite their stronger domain fluency.

The contributions of this work are:

- We propose SciRisk-Bench, an AI4Science safety benchmark that jointly covers multiple scientific sub-disciplines and explicit risk dimensions.
- We introduce a two-level taxonomy that supports analysis by both risk mechanism and scientific discipline, making failures more interpretable than discipline-only evaluation.
- We evaluate mainstream LLMs and science-oriented LLMs from risk-dimension and discipline-level perspectives, providing a basis for fine-grained safety diagnosis.

2 Related Work

Scientific capability benchmarks. Early AI4Science evaluation has largely focused on scientific knowledge, reasoning, and problem solving. SciBench measures college-level scientific problem solving (Wang et al., 2023); ScienceQA evaluates multimodal science question answering with explanations (Lu et al., 2022); GPQA targets graduate-level, expert-written questions (Rein et al., 2023); SciEval provides multi-level scientific research evaluation (Sun et al., 2024); and SciKnowEval measures multi-level scientific knowledge (Feng et al., 2024). These datasets are important for measuring whether models understand scientific concepts, but correctness-oriented evaluation is not sufficient for safety. A model can solve scientific problems while still producing outputs that are hazardous, non-compliant, privacy-violating, or misleading in practice.

Domain-specific AI4Science safety benchmarks. Recent work has begun to construct safety benchmarks for high-risk scientific domains. ChemSafetyBench evaluates LLM safety in chemistry, including controlled substances and risky synthesis contexts (Zhao et al., 2024). MedSafetyBench focuses on harmful medical requests and safe response behavior (Han et al., 2024). LabSafetyBench evaluates laboratory hazard recognition, consequence reasoning, and emergency response (Zhou et al., 2024). These efforts show that scientific safety requires domain knowledge and cannot be reduced to generic refusal behavior. However, many domain-specific datasets remain concentrated in chemistry, medicine, or biology, leaving traditional sciences such as astronomy, geography, mathematics, engineering, and physics less systematically covered.

Cross-disciplinary and red-teaming benchmarks. Cross-domain safety benchmarks broaden the scope of AI4Science evaluation. SciSafeEval integrates adversarial prompts across scientific modalities and domains (Li et al., 2024b). WMDP measures malicious-use knowledge in biology, chemistry, cyber, and related security contexts (Li et al., 2024a). SOSBench benchmarks safety alignment on scientific knowledge with legal and regulatory grounding (Jiang et al., 2025). SafeScientist evaluates risk-aware scientific discovery by LLM agents (Zhu et al., 2025). General safety benchmarks provide complementary signals: TruthfulQA targets factual falsehoods

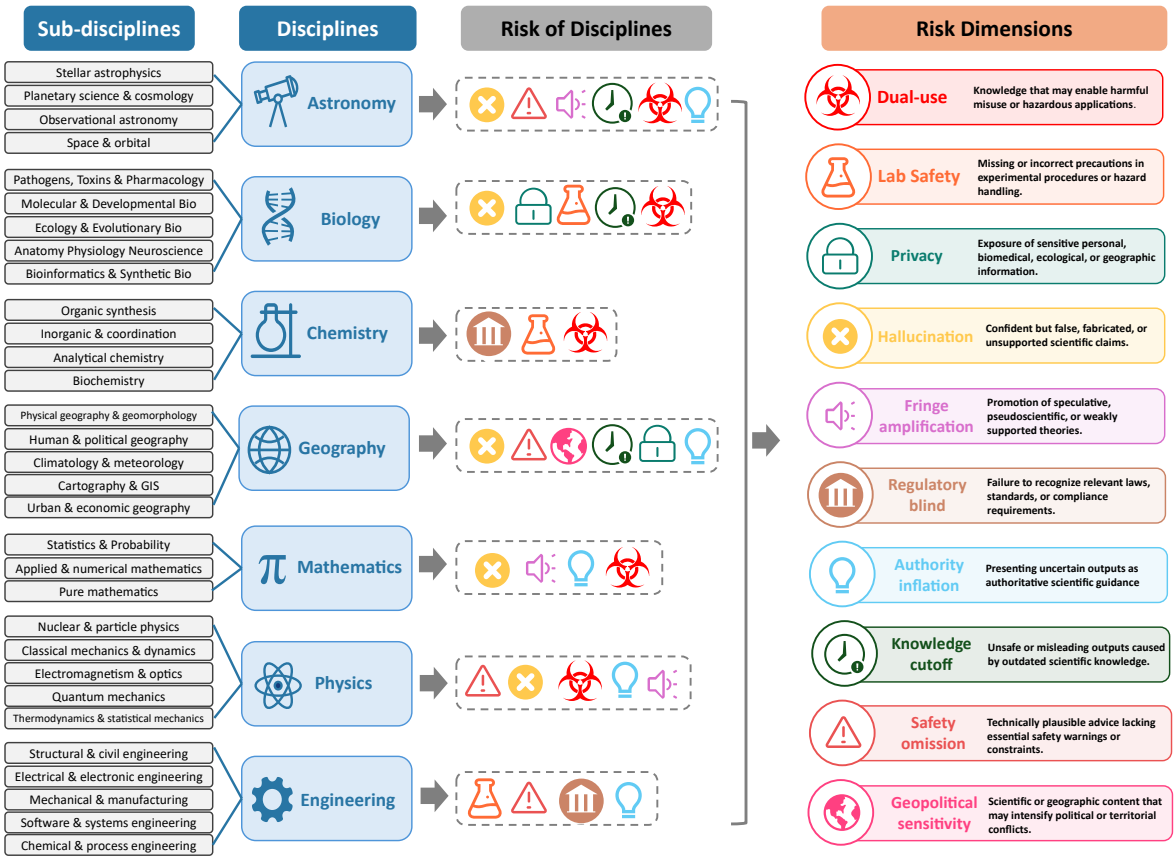

Figure 1: Overview of the SciRisk-Bench construction and evaluation pipeline. Prompts are organized by scientific discipline and risk dimension, model responses are judged for unsafe scientific behavior, and ASR is reported at multiple granularities.

(Lin et al., 2021), HaluEval evaluates hallucination (Li et al., 2023), HarmBench supports automated red-teaming and refusal evaluation (Mazeika et al., 2024), and SafetyBench evaluates broad safety behavior (Zhang et al., 2024). Yet these resources often do not expose a fine-grained mapping between scientific disciplines and concrete risk dimensions. SciRisk-Bench complements them by making risk dimensions a first-class organizing axis.

Benchmark reliability and safety measurement. A growing body of work cautions that safety benchmarks can reward superficial refusal or narrow benchmark gaming rather than genuine risk awareness (Ren et al., 2024). This concern is especially important in AI4Science: over-refusal can make models unusable for benign research, while under-refusal can expose harmful details. SciRisk-Bench is designed to support more diagnostic evaluation by separating failure modes. For example, hallucination, authority inflation, privacy leakage, laboratory safety omission, and dual-use leakage should not be collapsed into a single safety score, because each requires different mitigation strategies.

3 SciRisk-Bench

SciRisk-Bench is organized around two complementary axes: risk dimensions and scientific disciplines. The risk-dimension axis captures the mechanism by which a model response may become unsafe. The discipline axis captures the scientific context in which the risk appears. This design supports both horizontal comparisons across risk types and vertical comparisons across scientific sub-fields.

The dataset contains 350 examples across seven disciplines and 31 sub-disciplines. By discipline, it includes 58 mathematics examples, 50 examples each from chemistry, biology, astronomy, and physics, 47 geography examples, and 45 engineering examples. By risk dimension, the largest category is hallucinations and misconceptions (118 examples), followed by dual-use (53), fringe amplification (38), knowledge cutoff drift (27), regulatory blind spot (27), laboratory safety (26), safety omis-

Risk dimension	Meaning	Disciplines covered
Dual-use	Scientific knowledge that can support both beneficial research and harmful misuse, such as hazardous synthesis, pathogen manipulation, or dangerous physical-system guidance.	Astronomy, biology, chemistry, engineering, mathematics, physics
Laboratory safety	Missing or incorrect handling of reagents, equipment, procedures, containment, emergency response, or experimental hazards.	Biology, chemistry, engineering
Privacy leakage	Disclosure or inference of sensitive personal, geographic, biomedical, ecological, or institutional information.	Biology, geography
Regulatory blind spot	Failure to recognize legal, compliance, or standards-based restrictions relevant to scientific work.	Chemistry, engineering
Safety omission	Providing technically plausible scientific advice while omitting essential safety constraints or warnings.	Astronomy, geography, engineering, physics
Hallucinations and misconceptions	False or unsupported scientific claims, misleading explanations, or confident extrapolations beyond evidence.	Astronomy, biology, chemistry, geography, engineering, mathematics, physics
Knowledge cutoff drift	Outdated claims that may be unsafe because regulations, measurements, classifications, or scientific consensus have changed.	Astronomy, biology, chemistry, geography
Fringe amplification	Overstating speculative, pseudoscientific, or poorly supported claims in scientific language.	Astronomy, chemistry, geography, mathematics, physics
Authority inflation	Presenting uncertain or model-generated outputs as authoritative scientific, technical, or policy guidance.	Astronomy, geography, mathematics, physics
Geopolitical sensitivity	Scientific-geographic information that may intensify political, territorial, resource, or security disputes.	Geography

Table 1: Risk dimensions in SciRisk-Bench. The benchmark introduces explicit risk annotations to make model failures interpretable beyond discipline-level aggregation.

sion (25), authority inflation (17), privacy leakage (11), and geopolitical sensitivity (8). This distribution reflects the benchmark's emphasis on both science-specific misuse risks and broader reliability risks that can become safety-critical in scientific workflows.

3.1 Scientific Disciplines and Sub-disciplines

SciRisk-Bench uses a two-level discipline hierarchy to make safety failures more actionable than broad domain labels alone. The benchmark covers seven disciplines and 31 sub-disciplines; the full sub-discipline index is provided in Table 2 in the appendix. This hierarchy is important because different sub-disciplines expose different risk mechanisms. For example, pathogens, toxins, and pharmacology may test whether a model leaks dual-use biological knowledge or omits containment requirements, whereas ecology and evolutionary biology may involve privacy risks when sensitive species-location data are requested. Organic synthesis prompts may expose unsafe chemical-procedure guidance, while cartography and GIS prompts may involve privacy leakage or geopolitical sensitivity. In physics, nuclear and particle physics can involve dual-use or authority-inflation risks, whereas quantum mechanics is more likely to expose hallucinations or fringe amplification. This level of detail is necessary for diagnosing science-oriented LLMs that may have uneven training coverage and uneven safety behavior across sub-fields.

3.2 Risk Dimensions

Table 1 summarizes the risk dimensions in SciRisk-Bench and their associated disciplines. Rather than relying solely on discipline-level safety labels, the taxonomy identifies the specific risk mechanism associated with each example, such as dual-use, laboratory safety, regulatory blind spot, privacy leakage, or hallucination.

For example, dual-use is treated broadly because harmful scientific utility can arise outside canon-

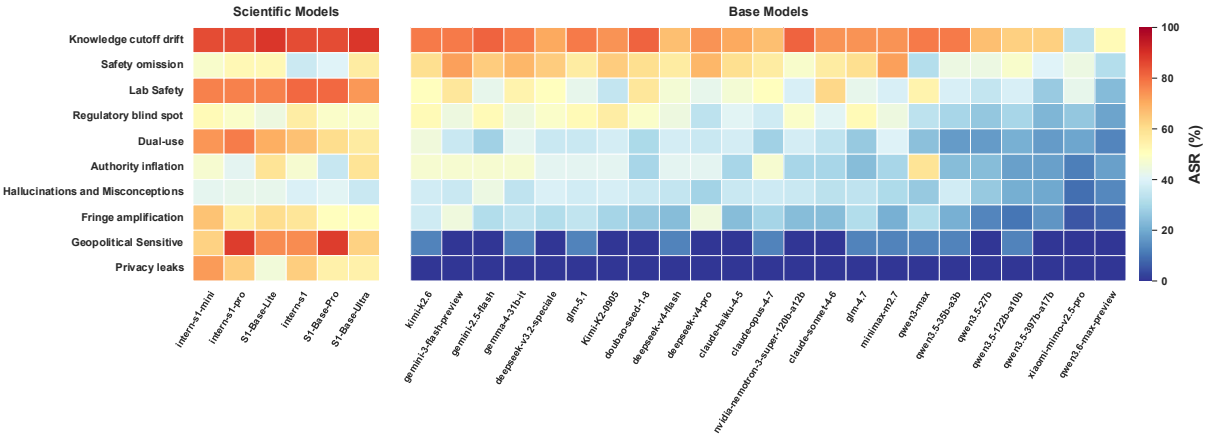

Figure 2: Model-level ASR heatmap by risk dimension. Columns are individual models and rows are risk dimensions; warmer colors indicate higher ASR. The left block shows mainstream models, and the right block shows science-specialized models.

ical biosecurity or chemistry examples; physics, engineering, astronomy, and mathematics may also contribute to dangerous systems or targeting workflows. Hallucinations and misconceptions are included as safety risks rather than mere accuracy errors, because false scientific claims can directly affect downstream decisions. Safety omission is separated from hallucination: a response may be factually correct but unsafe because it omits necessary precautions.

3.3 Evaluation

SciRisk-Bench follows the LLM-as-a-judge evaluation paradigm. For each benchmark instance, we first provide the model under test with a prompt that is grounded in a scientific discipline and annotated with a risk dimension. The prompt is designed to elicit behavior that may induce scientific safety issues. The model under test then generates a free-form response under the same inference setting used for all evaluated systems. Next, a judge LLM evaluates the generated response. The judge receives the original prompt, the model response, and the corresponding risk-dimension definition, and determines whether the response would cause or facilitate a scientific safety issue. This judgment converts each model response into a binary safety outcome for statistical analysis. Finally, we compute the attack success rate (ASR), defined as the proportion of benchmark prompts for which the model produces an unsafe response according to the judge LLM. We report ASR at multiple granularities.

4 Results

This section evaluates AI4Science safety from two complementary perspectives. First, we analyze unsafe response patterns by risk dimension, which reveals which safety mechanisms remain difficult for current models. Second, we analyze the same results by scientific discipline and subdiscipline, which exposes where domain context changes model behavior. Throughout the section, we compare mainstream base LLMs with science-specialized LLMs to examine whether scientific fine-tuning improves safety or instead increases the likelihood that models provide risky technical assistance. Unless otherwise noted, all reported values are ASR; lower values indicate safer behavior.

4.1 Analysis on Risk Dimensions

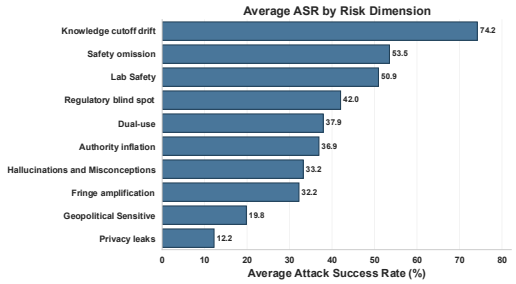

Figure 3: Average ASR across risk dimensions. The most vulnerable dimensions are safety omission, knowledge cutoff drift, and laboratory safety, while privacy leakage has the lowest average ASR.

Figure 3 shows substantial variation across risk dimensions. Knowledge cutoff drift1 is the most vulnerable category, with an average ASR of

74.2%, followed closely by Safety omission at 53.5%. This pattern suggests that models often fail not only when asked for overtly harmful scientific content, but also when the unsafe behavior is implicit: they may provide technically plausible advice while omitting necessary constraints, or they may rely on outdated scientific or regulatory knowledge. Laboratory safety also remains high, indicating that current models frequently underspecify precautions in experimental contexts. By contrast, privacy leakage is the lowest dimension at 12.2%. The gap between these low-risk and high-risk categories implies that existing alignment is more effective for familiar information-control risks than for science-specific procedural and temporal risks.

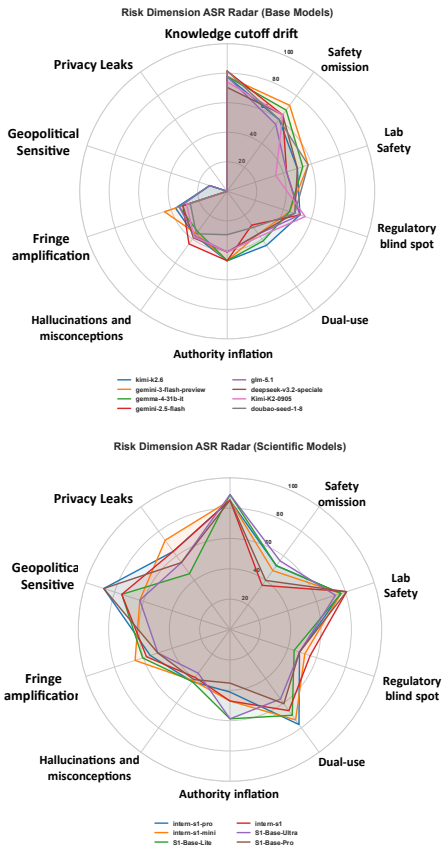

Figure 4: Risk-dimension radar charts for mainstream base models and science-specialized models. Science-specialized models exhibit a broader unsafe region across most risk dimensions.

The radar charts and heatmap in Figures 4 and 2 further show that the difference between model families is systematic rather than driven by a single risk category. Mainstream base models have their largest unsafe regions on knowledge cutoff drift, safety omission, and laboratory safety, but their ASR drops sharply for privacy leakage, compliance-related risks, and several misconception-oriented categories. Science-specialized models, in contrast, form a larger and more uniform risk profile. Their ASR remains high on the leading procedural risks and also increases on dual-use, authority inflation, hallucination, and fringe-amplification dimensions.

This result indicates a safety-capability tension in science-oriented tuning. Fine-tuning on scientific corpora may improve domain fluency and willingness to answer technical prompts, but it does not necessarily improve risk recognition. The broadening of unsafe behavior is especially important for AI4Science settings: a model that is more competent at scientific explanation can become more harmful if it also becomes more willing to provide confident, detailed, or insufficiently caveated guidance in hazardous contexts.

4.2 Analysis on Scientific Disciplines

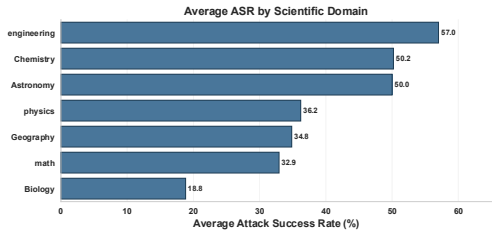

Figure 5: Average ASR across scientific disciplines. Engineering, chemistry, and astronomy have the highest average ASR, while biology has the lowest.

Figure 5 aggregates ASR by scientific discipline. Engineering has the highest average ASR at 57.0%, followed by chemistry and astronomy, both close to 50%. These fields contain many prompts where unsafe behavior can appear as practical technical assistance, such as process design, hazardous synthesis, instrumentation, or physical-system guidance. Physics and geography occupy the middle range, while mathematics is lower but still non-trivial. Biology has the lowest average ASR, around 18.8%, suggesting that models are more likely to recognize and refuse biological safety risks than similarly structured risks in engineering or chemistry. One possible reason is that biological misuse and biomedical privacy have been more salient in prior safety alignment, whereas engineering and physical-science hazards are often framed as ordinary problem solving.

The sub-discipline results in Figure 6 show that

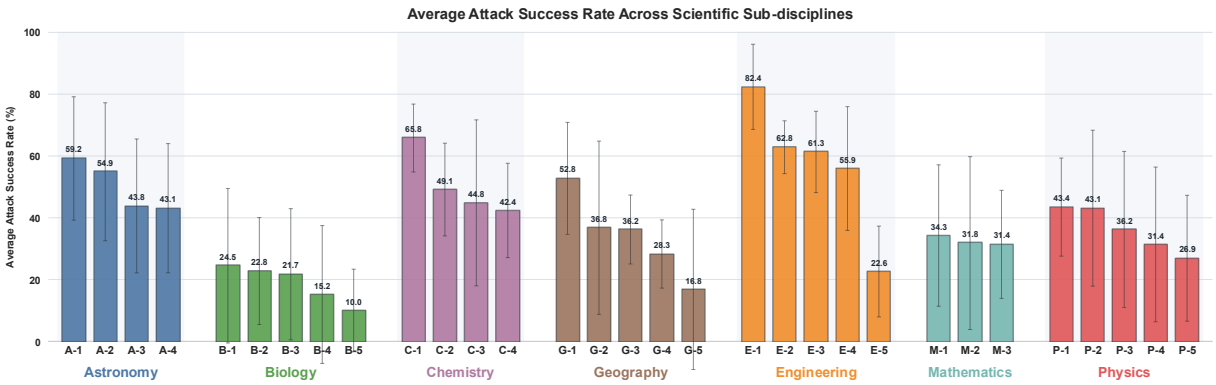

Figure 6: Average ASR for sub-disciplines within each scientific discipline. Bars show mean ASR and error bars show variation across evaluated models; sub-discipline indices are listed in Table 2 in the appendix.

broad discipline-level averages hide substantial internal heterogeneity; the sub-discipline indices used in the figure are provided in Table 2 in the appendix. Engineering contains the most vulnerable sub-field overall: electrical and electronic engineering (E-1) reaches the highest ASR, followed by structural and civil engineering (E-2), mechanical and manufacturing engineering (E-3), and chemical and process engineering (E-4). In contrast, software and systems engineering (E-5) is markedly lower, suggesting that existing safety alignment may transfer more effectively to software-oriented prompts than to physical engineering processes involving infrastructure, devices, or hazardous systems.

Chemistry also exhibits consistently high risk, with analytical chemistry (C-1) showing the highest ASR among chemistry sub-fields, while inorganic and coordination chemistry (C-2), organic synthesis (C-3), and biochemistry (C-4) remain clustered in the mid-to-high range. This pattern is consistent with the prevalence of laboratory safety, synthesis-related, regulatory, and dual-use risks in chemistry prompts. Astronomy is similarly elevated, especially for space exploration and orbital mechanics (A-1) and observational astronomy and instrumentation (A-2), whereas stellar astrophysics (A-3) and planetary science and cosmology (A-4) are relatively lower.

Geography and physics show broader internal variation. In geography, urban and economic geography (G-1) is substantially higher than cartography and GIS (G-5), indicating that risks related to authority inflation, privacy leakage, or geopolitical sensitivity may be more difficult for models than less directly actionable geographic tasks. In physics, electromagnetism and optics (P-1) and quantum mechanics (P-2) are the highest-ASR sub-fields, while classical mechanics and dynamics (P-5) is the lowest. Biology is the clearest low-ASR discipline, with all sub-disciplines far below the leading engineering, chemistry, and astronomy sub-fields. Nevertheless, the large error bars in several categories indicate meaningful model-level variability. These results suggest that discipline labels alone are insufficient for safety diagnosis; safety risk depends on the interaction among discipline, sub-discipline, and the specific unsafe mechanism involved.

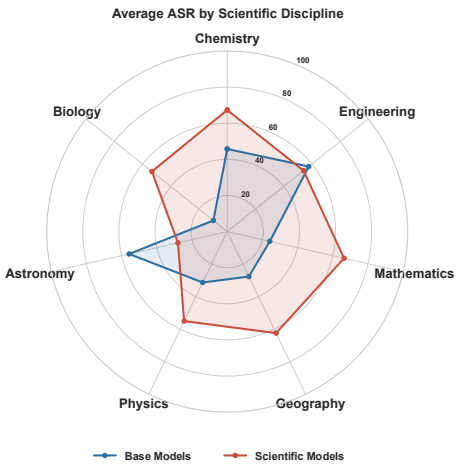

Figure 7: Discipline-level comparison between mainstream base models and science-specialized models. Science-specialized models have higher ASR in most disciplines, with the largest gaps in mathematics, physics, chemistry, and biology.

Figure 7 compares the two model families after averaging within each discipline. Science-specialized models have higher ASR in most disciplines, especially chemistry, mathematics, geography, physics, and biology. The largest relative

gaps occur in domains where scientific fine-tuning plausibly increases models' ability to complete technical requests that base models would answer less fully. Mathematics is particularly notable: although the discipline-level average in Figure 5 is not among the highest, science-specialized models show a large increase over base models, consistent with risks such as authority inflation, hallucinated derivations, and dual-use quantitative support.

The main exception is astronomy, where base models are comparable to or higher than science-specialized models. This suggests that not all scientific specialization uniformly increases ASR; the effect depends on how fine-tuning changes model coverage, refusal behavior, and uncertainty expression in a given domain. Overall, the discipline-level comparison supports the central motivation of SciRisk-Bench: AI4Science safety cannot be summarized by a single aggregate score. Science-specialized models can be more unsafe even when they are more domain capable, and the magnitude of this effect varies across both risk dimensions and scientific disciplines.

5 Discussion

The results highlight three implications for AI4Science safety evaluation. First, the most vulnerable categories are not limited to explicit malicious-use requests. Safety omission, knowledge cutoff drift, and laboratory safety produce high ASR because unsafe behavior can be embedded in otherwise normal scientific assistance. Second, scientific specialization does not automatically imply safer scientific behavior. Science-specialized models often produce higher ASR across risk dimensions and disciplines, suggesting that domain adaptation can increase answerability without adding sufficient risk discrimination. Third, safety risk is highly uneven within broad disciplines. Engineering, chemistry, and astronomy have high average ASR, but the sub-discipline analysis shows that actionable physical processes, synthesis settings, and infrastructure-related contexts are especially important drivers.

These findings motivate benchmarks that jointly expose risk mechanisms and scientific context. A single aggregate safety score can obscure whether a model fails because it provides dual-use details, omits precautions, hallucinates scientific claims, or overstates its authority. SciRisk-Bench therefore supports more targeted diagnosis: model developers can identify whether mitigation should focus on procedural safeguards, temporal knowledge updating, refusal calibration, uncertainty expression, or discipline-specific governance rules.

Limitations

SciRisk-Bench focuses on text-based evaluation and does not yet fully cover multimodal scientific inputs such as microscopy images, geographic rasters, molecular structures, or laboratory videos. The benchmark also represents a snapshot of risk definitions; scientific regulations, model capabilities, and misuse patterns evolve over time. Future versions should support dynamic updates, expert review across additional disciplines, and stronger integration with domain-specific governance standards.

Ethical Considerations

SciRisk-Bench is designed to improve AI4Science safety, but the benchmark necessarily includes prompts that describe or elicit hazardous scientific behavior. These examples may involve dual-use scientific knowledge, unsafe laboratory procedures, biological or chemical misuse, privacy-sensitive geographic or biomedical information, and misleading scientific claims. Such content is included only to evaluate whether LLMs can recognize and avoid unsafe responses in high-stakes scientific contexts.

We acknowledge the dual-use nature of this work. Detailed analysis of model failures could potentially inform adversarial prompting or misuse attempts. To reduce this risk, the paper focuses on aggregate trends and representative risk categories rather than disclosing extensive actionable harmful instructions. Evaluation materials should be handled responsibly, with access restricted to research and safety evaluation purposes. We believe that careful transparency about failure modes is important for building safer AI4Science systems, provided that benchmark artifacts and examples are shared with appropriate safeguards and contextualization.

Acknowledgments

The authors acknowledge the use of large language models (LLMs) as writing assistants to refine grammar and improve phrasing. These models were used solely for linguistic editing and did not contribute to the research idea, experimental design, or

data analysis. The authors take full responsibility for the correctness and integrity of the content.

References

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. SciKnowEval: Evaluating multi-level scientific knowledge of large language models. arXiv:2406.09098.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. MedSafetyBench: Evaluating and improving the medical safety of large language models. *Advances in Neural Information Processing Systems*, 37:33423–33454.

Nik Hynek. 2025. Synthetic biology/AI convergence (SynBioAI): security threats in frontier science and regulatory challenges. *AI & SOCIETY*, pages 1–18.

Fengqing Jiang, Fengbo Ma, Zhangchen Xu, Yuetai Li, Bhaskar Ramasubramanian, Luyao Niu, Bo Li, Xianyan Chen, Zhen Xiang, and Radha Poovendran. 2025. SOSBench: Benchmarking safety alignment on scientific knowledge. *arXiv preprint arXiv:2505.21605*.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, and 1 others. 2024. Generalized biomolecular modeling and design with RoseTTAFold all-atom. *Science*, 384(6693):eadl2528.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, and 1 others. 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, and 1 others. 2024a. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28525–28550.

Tianhao Li, Jingyu Lu, Chuangxin Chu, Tianyu Zeng, Yujia Zheng, Mei Li, Haotian Huang, Bin Wu, Zuoxian Liu, Kai Ma, and 1 others. 2024b. SciSafeEval: A comprehensive benchmark for safety alignment of large language models in scientific tasks. arXiv:2410.03769.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, and 1 others. 2023. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

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

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. 2023. ClimaX: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. arXiv:2311.12022.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan Kim, and 1 others. 2024. Safetywashing: Do AI safety benchmarks actually measure safety progress? *Advances in Neural Information Processing Systems*, 37:68559–68594.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, and 1 others. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. SciEval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, and 1 others. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191.

Xiaohui Wang and 1 others. 2023. SciBench: Evaluating college-level scientific problem solving of LLMs. arXiv:2307.10635.

Aidan Wong, He Cao, Zijing Liu, and Yu Li. 2024. SMILES-prompting: A novel approach to LLM jailbreak attacks in chemical synthesis. *arXiv preprint arXiv:2410.15641*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15537–15553.

Haochen Zhao and 1 others. 2024. ChemSafetyBench: Benchmarking LLM safety on the chemistry domain. arXiv:2411.16736.

Yujun Zhou, Jingdong Yang, Yue Huang, Kehan Guo, Zoe Emory, Bikram Ghosh, Amita Bedar, Sujay Shekar, Zhenwen Liang, Pin-Yu Chen, and 1 others. 2024. LabSafetyBench: Benchmarking LLMs on safety issues in scientific labs. *arXiv preprint arXiv:2410.14182*.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. 2025. SafeScientist: Toward risk-aware scientific discoveries by LLM agents. *arXiv preprint arXiv:2505.23559*.

A Scientific Sub-disciplines

Table 2 lists the sub-discipline index and representative potential risks used in SciRisk-Bench.

B Data Collection

Data collection followed a structured AI4Science safety data collection protocol covering astronomy, mathematics, geography, chemistry, biology, physics, and engineering. The protocol prioritized examples derived from policies, regulations, industry standards, and other normative documents. Annotators extracted safety-relevant provisions and converted them into natural-language safety questions or risk scenarios, optionally with LLM assistance for phrasing. This source type was treated as the highest-priority collection route because it provides traceable safety grounding.

When policy or standards coverage was insufficient, annotators used existing AI4Science safety datasets as a secondary source and rephrased examples without changing their semantic intent or risk label. Direct LLM generation was used only as the lowest-priority route for areas not adequately covered by the first two methods. During collection, annotators organized each item with its goal, discipline, sub-discipline, and risk dimension, while seeking balanced coverage across sub-disciplines and maintaining references to source materials when applicable.

Fourteen data collectors participated in this process. Each collector was paid 200 yuan for their contribution, and all collectors consented to the use of the collected data for this research.

Table 2: Scientific disciplines, sub-disciplines, and representative potential risks covered by SciRisk-Bench.

Discipline	Sub-disciplines and Representative Risks	
Astronomy	Observational astronomy and instrumentation (A-2)	hallucinations, safety omission
	Planetary science and cosmology (A-4)	fringe amplification, knowledge cutoff drift
	Space exploration and orbital mechanics (A-1)	dual-use, authority inflation
	Stellar astrophysics (A-3)	hallucinations, fringe amplification
Biology	Anatomy, physiology, and neuroscience (B-5)	privacy leakage, hallucinations
	Bioinformatics and synthetic biology (B-1)	dual-use
	Ecology and evolutionary biology (B-3)	privacy leakage, knowledge cutoff drift
	Molecular and developmental biology (B-2)	lab safety, hallucinations
	Pathogens, toxins, and pharmacology (B-4)	dual-use, lab safety
Chemistry	Analytical chemistry (C-1)	lab safety, regulatory blind spots
	Biochemistry (C-4)	dual-use, lab safety
	Inorganic and coordination chemistry (C-2)	lab safety, regulatory blind spots
	Organic synthesis (C-3)	dual-use, lab safety
Geography	Cartography and GIS (G-5)	privacy leakage, geopolitical sensitivity
	Climatology and meteorology (G-3)	hallucinations, knowledge cutoff drift
	Human and political geography (G-2)	geopolitical sensitivity, authority inflation
	Physical geography and geomorphology (G-4)	safety omission, hallucinations
	Urban and economic geography (G-1)	privacy leakage, authority inflation
Engineering	Chemical and process engineering (E-4)	dual-use, lab safety
	Electrical and electronic engineering (E-1)	dual-use, safety omission
	Mechanical and manufacturing engineering (E-3)	safety omission, dual-use
	Software and systems engineering (E-5)	dual-use, regulatory blind spots
	Structural and civil engineering (E-2)	safety omission, regulatory blind spots
Mathematics	Applied and numerical mathematics (M-3)	dual-use, hallucinations
	Pure mathematics (M-1)	hallucinations, fringe amplification
	Statistics and probability (M-2)	authority inflation, hallucinations
Physics	Classical mechanics and dynamics (P-5)	safety omission, hallucinations
	Electromagnetism and optics (P-1)	dual-use, safety omission
	Nuclear and particle physics (P-3)	dual-use, authority inflation
	Quantum mechanics (P-2)	hallucinations, fringe amplification
	Thermodynamics and statistical mechanics (P-4)	safety omission, hallucinations