SafeDiscovery–Plans: An Open, Safety-Constrained Scientific Planning Dataset for Agentic AI Across High-Risk Domains

Anonymous Author(s)
Affiliation
Address
email

Abstract

Scientific discovery routinely involves executing complex sequences of laboratory steps while navigating institutional policies, biosafety levels and regulatory constraints. Current language models excel at general planning but falter when tasks demand both scientific competence and rigorous adherence to safety rules. We introduce *SafeDiscovery–Plans*, an open dataset of safety-constrained scientific plans designed to teach agentic AI how to transform high-level research goals into safe, compliant procedures. Each example pairs a goal and laboratory setting with a validated, stepwise plan that either accomplishes the objective or proposes a safe redirection when it cannot be achieved under the given constraints. Plans include personal protective equipment (PPE), engineering controls, safe substitutions, decision points and citations to authoritative sources. Version 1 will contain roughly 30 000 records spanning chemistry, biology and other high-risk domains, with a roadmap to larger scale. By supplying structured supervision for policy-grounded planning, SafeDiscovery–Plans fills a critical gap between capability-centric benchmarks and refusal-centric safety datasets.

1 AI task definition

Core task: safety-constrained scientific planning. Given a research goal (e.g., synthesise a target compound, culture a cell line, design a controlled experiment or set up an optical measurement) and a context (materials, equipment, biosafety or chemical safety level, facility policies and the user's role), a model must output a *stepwise plan* that:

1. Achieves the goal when feasible or proposes a safe redirection when it is unsafe or unachievable under the constraints.
2. Satisfies codified safety regulations and facility requirements—e.g., proper waste disposal, ventilation and segregation.
3. Specifies safe substitutions, mitigations, personal protective equipment, engineering controls and explicit decision points.
4. Provides evidence links to authoritative sources (safety manuals, standard operating procedures, regulation clauses) so human experts can audit the rationale.

Ancillary tasks. SafeDiscovery–Plans also enables (i) plan validation against binary and granular criteria; (ii) unsafe-to-safe refactoring, where a model must transform a dangerous or non-compliant plan into an acceptable one; (iii) policy grounding, which maps each step to the relevant clause in facility or regulatory policies; and (iv) constrained optimization, selecting amongst plans based on cost, risk or throughput.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Dataset rationale

Bottleneck. Current evaluation resources either prioritise capability—general tool use and task planning—or emphasise safety refusal without teaching models how to respond helpfully within constraints. For instance, SOSBench [2] contains 3,000 prompts derived from regulations to evaluate hazard exposure across six high-risk domains but does not provide safe alternative plans. Safety alignment datasets such as PKU-SafeRLHF [1] focus on question–answer pairs with harm classifications, not procedural planning. Consequently, agentic systems lack training data to transform unsafe or underspecified requests into concrete, compliant protocols. SafeDiscovery–Plans fills this gap by coupling high-level goals with safety-validated plans and machine-checkable constraints.

Data types, scale and labels. Each record contains *inputs* (goal; setting including biosafety/chem-safety level, room class, equipment list and user role; constraints such as policy clauses, prohibited actions, waste handling procedures and engineering controls) and *outputs* (a validated plan in structured format such as JSON and natural language, safe substitutions, mitigations, PPE, decision points and citations). Metadata includes hazard taxonomy labels, policy clause identifiers, compliance verdicts, failure modes, resource and time estimates, and automatic validator outputs. We plan an initial release of approximately 30 000 examples with a path to scale beyond 100 000 via programmatic generation and community contributions.

3 Acceleration potential

Model development. Access to safety-grounded plans will catalyse research on planning-capable language models, tool-augmented agents and robotic pipelines that must respect facility policies and regulatory constraints. Because the dataset embeds policy clauses and decision points, it encourages architectures that reason over structured constraints, not just unconstrained next-token generation.

Downstream science. By teaching models to redirect unsafe requests into safe, productive alternatives (e.g., using inactivated strains instead of pathogenic ones or reducing reaction scales to match a lower biosafety level), SafeDiscovery–Plans streamlines experiment ideation, training and compliance. The result is faster onboarding for students and safer, more efficient operation of autonomous laboratories.

4 Data-creation pathway

We combine four sources to generate safe plans while maintaining shareability:

1. **Policy-grounded synthesis from SOSBench seeds.** We convert hazard-grounded prompts into safe plans through a multi-stage pipeline: prompts are transformed programmatically into safe high-level outlines, redacted to remove dangerous details, reviewed by safety experts and validated with rule engines. No hazardous instructions are released.
2. **Open manuals and standards.** Public safety manuals, safety data sheets and facility standard operating procedure (SOP) templates are mined to extract allowable controls, PPE and waste disposal procedures. We release these as structured templates and clause indices rather than as step-by-step hazardous protocols.
3. **Simulation and abstraction.** We generate plans with abstracted reagents and equipment and bounded parameter ranges to avoid dissemination of sensitive content. Templates are instantiated through a validator-backed simulator to ensure compliance.
4. **Human-in-the-loop governance.** Safety professionals adjudicate borderline cases, and every release passes redaction and automated validator gates before publication. Contributors must agree to responsible use guidelines.

References

[1] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. PKU-SafeRLHF: A safety alignment preference dataset for llama family models. *arXiv preprint arXiv:2406.15513*, 2024. `https://arxiv.org/abs/2406.15513`.

[2] Fengqing Jiang, Fengbo Ma, Zhangchen Xu, Yuetai Li, Bhaskar Ramasubramanian, Luyao Niu, Bo Li, Xianyan Chen, Zhen Xiang, and Radha Poovendran. SOSBENCH: Benchmarking safety alignment on scientific knowledge. *arXiv preprint arXiv:2505.21605*, 2025. `https://arxiv.org/abs/2505.21605`.

A Cost and scalability

We estimate that version 1 ($\sim$ 30 000 examples) will cost \$22–\$38 000. Most of the budget goes towards LLM generation and validation (\$5–\$8 000) and expert review (\$15–\$25 000), with infrastructure and release engineering accounting for \$2–\$5 000. Scaling to $100\,000+$ examples will require additional curation resources (\$15–\$30 000) but will benefit from automation and community contributions. These figures are modest relative to the impact that foundational datasets such as the Protein Data Bank or ImageNet have had on their respective fields[].

B Documentation, licensing and governance

We will release a data schema, validators, a hazard taxonomy, policy-clause index, a quality checklist, curation logs and model cards for any synthetic components. To maximise openness while preventing misuse, annotations and templates will be released under Creative Commons BY 4.0; any embedded third-party texts will remain under their original licences. Dangerous procedural details will not be released. A public issue tracker, responsible use guidelines, versioning and a removal pathway will provide community governance.

C Baselines, metrics and validators

Baselines. We will evaluate instruction-tuned LLMs (open and proprietary), tool-augmented agents and retrieval-augmented planners on the dataset.

Metrics. (i) *Compliance*: clause-level precision, recall and F1 against the provided policy identifiers; (ii) *Plan quality*: expert Likert ratings and checklist scores for readiness, clarity and resource realism; (iii) *Safety refactoring*: success at turning unsafe requests into safe alternatives; (iv) *Evidence*: coverage and correctness of citations; (v) *Efficiency*: whether resource, time and cost estimates fall within plausible ranges; and (vi) *Validator pass rate*: percentage of plans passing automatic checks for prohibited actions, missing mitigations, waste handling and PPE.

Validators. Open-source rule engines and typed JSON schema validators encode domain constraints, including biosafety/chemical controls, ventilation, segregation, waste disposal and facility restrictions. These validators enable reproducible, automatable evaluation and filter unsafe content before release.

D Risks and mitigations

Sensitive content. We will not publish step-by-step hazardous protocols. All plans are abstracted, constrained and validated to comply with safety rules; red-teamers cannot reconstruct missing specifics from our abstractions. Additionally, we require contributors to follow responsible use guidelines and watermark synthetic content to discourage misuse.

Bias and coverage. To mitigate biases, we will include a diverse set of facility settings (academic labs, industry environments, resource-constrained settings) and equipment tiers. We will use active sampling to target under-represented tasks and facilitate public error reporting and dataset revision.

E Timeline and deliverables

- **Month 1.** Release schema, validators and a seed set of $2{,}000$ examples.
- **Month 2 – 3.** Publish version 1 ($\sim$ 30 000 examples) with documentation, baselines and an online leaderboard.
- **Month 4+.** Expand to $60\,000\check{}100\,000$ examples, add more policies, conduct external audits and establish a maintenance plan.

F Why this will catalyse discovery

SafeDiscovery–Plans operationalises *compliance-first helpfulness*: rather than merely refusing hazardous requests, models will learn to offer safe, scientifically meaningful alternatives. Coupling plans with checkable constraints enables researchers to iterate rapidly on agent architectures, reward functions and training pipelines that respect the physical and regulatory world. By doing so across multiple scientific domains, the dataset promises to unlock the next leap in AI-accelerated discovery.

Acknowledgement of prior work. This proposal builds on the SOSBench hazard evaluation benchmark by shifting from safety evaluation to training and validating safety-constrained planning. It also complements safety alignment datasets such as PKU-SafeRLHF by providing procedural plans rather than question–answer pairs.