# DATASHEET:
# SciRecipe

## MOTIVATION

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The SciRecipe dataset was created to enable language models to understand and generate experimental protocols in life science research. Existing datasets contain scientific text but do not cover the procedural and action centered characteristics of real laboratory workflows. This gap limits the development of models that need to reason about materials, equipment, stepwise operations, parameters, and execution logic. SciRecipe addresses this gap by providing structured protocol data and a diverse collection of tasks that support both understanding and practical problem solving in experimental settings.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
Anonymous during peer review.

**What support was needed to make this dataset?** (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
The creation of the dataset relied on institutional support for data collection, computing resources, and expert review.

**Any other comments?**
SciRecipe aims to promote reproducible research in scientific protocol generation and will be made available on HuggingFace Datasets for the research community.

## COMPOSITION

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each instance represents a scientific experimental protocol and a corresponding task specific question answer pair. All instances are text based. There is only one type of instance drawn from structured protocol documents.

**How many instances are there in total (of each type, if appropriate)?**
There are about twelve thousand structured protocols and over twenty thousand task specific question answer pairs.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset is a curated sample from a much larger collection of publicly available experimental protocols. It is designed to be diverse across scientific subfields rather than exhaustive. Representativeness was ensured by filtering for domain coverage and removing redundant entries.

**What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance contains cleaned protocol text a structured representation of its components and a task specific question answer pair generated from that protocol.

**Is there a label or target associated with each instance?** If so, please provide a description.
Yes. For each question there is an expected answer that serves as the target.E

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No. All required fields are present after data cleaning.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
No. Instances are independent.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind

them.
Yes. There are training and testing splits. The test split includes SciRecipe-Eval with controlled difficulty levels.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
Minor noise may exist due to original protocol variability but extensive cleaning and multi stage quality control were applied to minimize this.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self contained. It does not rely on external links after preprocessing. Source documents were processed offline and are not required for downstream use.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.
No. All data originate from publicly available scientific protocols.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
No.

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
No.

**Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
No.

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.
No.

**Any other comments?**
No.

## COLLECTION

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data were acquired from publicly available scientific protocol documents. These documents were extracted as raw text and then cleaned and structured. All task specific question answer pairs were derived from these processed protocols. The derived data were validated through multi stage automated checks and manual expert review.

**Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The source protocols were collected within a recent multi year window and represent long standing scientific literature. The structured dataset and its task instances were created during the preparation of this submission. The dataset will be first published after acceptance.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?
Collection used software extraction tools for PDF text processing manual curation and model assisted restructuring. Validation was carried out through consistency checks model based verification and expert inspection.

**What was the resource cost of collecting the data?** (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[1] for approaches in this area.)
The collection required standard computing resources for text extraction embedding based similarity search and

large scale data cleaning. The overall cost was modest and comparable to typical academic data preprocessing workloads. No unusual energy consumption was involved.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
Sampling was deterministic. Protocols were filtered by quality structural clarity diversity across scientific subfields and redundancy removal.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Anonymous domain experts were involved in reviewing a subset of instances. Their contribution was part of their regular research activity and no separate compensation applies.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
No formal review was required because the dataset contains only publicly available scientific documents and involves no human subject data.

**Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.
No.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
Not applicable.

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Not applicable.

**Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Not applicable.

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
Not applicable.

**Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Not applicable because the dataset does not contain data from individuals.

**Any other comments?**
None.

## PREPROCESSING / CLEANING / LABELING

**Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Yes. Extensive preprocessing and cleaning were performed. Raw scientific protocols were extracted from PDF files. Non textual elements were removed. Redundant and highly similar protocols were filtered through embedding based similarity checks. Protocols were restructured into a unified format through rule based parsing and model assisted summarization. Task specific question answer pairs were generated and validated through automated format checks and manual review.

**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
No. Only the cleaned and structured versions are retained because the raw PDFs are publicly available from their original sources and cannot be redistributed directly.

**Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.
The preprocessing relied on standard open source tools such as PDF text extraction software and embedding based similarity search. Additional internal scripts will be released with the dataset where possible.

**Any other comments?**
None.

## USES

**Has the dataset been used for any tasks already?** If so, please provide a description.

Yes. It has been used for training a model in this submission that focuses on experimental protocol understanding and generation.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
No. The dataset will be released on HuggingFace Datasets after acceptance.

**What (other) tasks could the dataset be used for?**
The dataset can support tasks such as protocol comprehension protocol transformation experimental planning troubleshooting parameter extraction safety analysis and other scientific reasoning tasks.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No. The dataset contains only publicly available scientific protocol text and does not involve human subject data. There are no foreseeable risks for unfair or harmful outcomes.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.
No.

**Any other comments?**
None.

## DISTRIBUTION

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
Yes. The dataset will be made publicly available to the research community.

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
The dataset will be distributed through HuggingFace Datasets. A DOI will be assigned after acceptance.

**When will the dataset be distributed?**
The dataset will be released after the completion of the peer review process.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
Yes. The dataset will be released under a permissive open license that will allow academic and non commercial use. The exact license will be specified upon release.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No. All source materials come from publicly accessible scientific documents.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

**Any other comments?**
None.

## MAINTENANCE

**Who is supporting/hosting/maintaining the dataset?**
The dataset will be maintained by the authors of this submission who remain anonymous during peer review.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
Contact information will be provided after acceptance to preserve anonymity during review.

**Is there an erratum?** If so, please provide a link or other access point.
No.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
Yes. Updates may be released to correct errors or add new instances. Updates will be announced through the HuggingFace dataset page.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time**

**and then deleted)?** If so, please describe these limits and explain how they will be enforced.
Not applicable.

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Older versions may be kept for a limited time if required for reproducibility. Users will be informed on the dataset page when a version becomes outdated.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Yes. Users may open issues or pull requests on the dataset repository after its release. Contributions will be reviewed before integration.

**Any other comments?**
None.

## REFERENCES

[1] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.