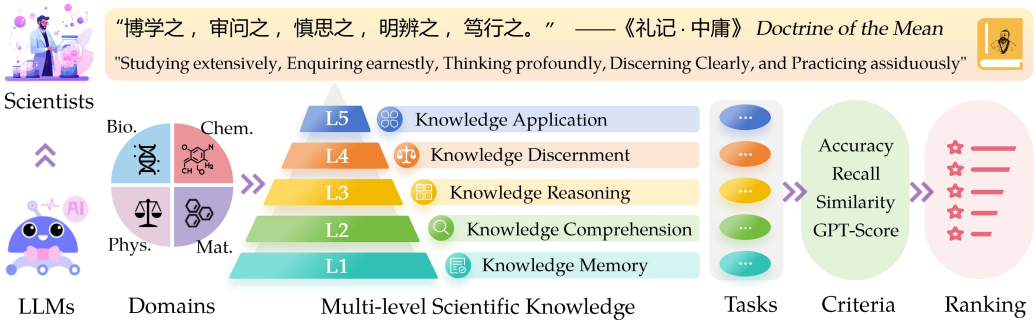
SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models

博学之， 审问之， 慎思之， 明辨之， 笃行之。

—— 《礼记 · 中庸》 *Doctrine of the Mean*



The **Sci**entific **Know**ledge **Eval**uation (**SciKnowEval**) benchmark for Large Language Models (LLMs) is inspired by the profound principles outlined in the "*Doctrine of the Mean*" from ancient Chinese philosophy. This benchmark is designed to assess LLMs based on their proficiency in **Studying Extensively**, **Enquiring Earnestly**, **Thinking Profoundly**, **Discerning Clearly**, and **Practicing Assiduously**. Each of these dimensions offers a unique perspective on evaluating the capabilities of LLMs in handling scientific knowledge.

News

- **[Sep 2024]** We released an evaluation report of OpenAI o1 with SciKnowEval.
- **[Sep 2024]** We have updated the SciKnowEval paper in arXiv.
- **[Jul 2024]** We have recently added the Physics and Materials to SciKnowEval. You can access the dataset here and check out the leaderboard here.
- **[Jun 2024]** We released the SciKnowEval Dataset and Leaderboard for Biology and Chemistry.

Table of Contents

Overview

Evaluated Abilities

- 📖 **L1**: **Studying extensively** (i.e., *knowledge memory*). This dimension evaluates the breadth of an LLM's knowledge across various scientific domains. It measures the model's ability to remember a wide range of scientific concepts.
- ❓ **L2**: **Enquiring earnestly** (i.e., *knowledge comprehension*). This aspect focuses on the LLM's capacity for deep enquiry and exploration within scientific contexts, such as analyzing scientific texts, identifying key concepts, and questioning relevant information.
- 💡 **L3**: **Thinking profoundly** (i.e., *knowledge reasoning*). This criterion examines the model's capacity for critical thinking, logical deduction, numerical calculation, function prediction, and the ability to engage in reflective reasoning to solve problems.
- 🔨 **L4**: **Discerning clearly** (i.e., *knowledge discernment*). This aspect evaluates the LLM's ability to make correct, secure, and ethical decisions based on scientific knowledge, including assessing the harmfulness and toxicity of information, and understanding the ethical implications and safety concerns related to scientific endeavors.
- 🔬 **L5**: **Practicing assiduously** (i.e., *knowledge application*). The final dimension assesses the LLM's capability to apply scientific knowledge effectively in real-world scenarios, such as analyzing complex scientific problems and creating innovative solutions.

Domains and Tasks

Domain	Ability	Task Name	Task Type	Data Source	Method	#Questions
Bio.	L1	Biological Literature QA	MCQ	Literature Corpus	I	14,862
		Protein Property Identification	MCQ	UniProtKB	III	1,500
		Protein Captioning	GEN	UniProtKB	III	930
	L2	Drug-Drug Relation Extraction	RE	Bohrium	II	464
		Biomedical Judgment and Interpretation	T/F	PubMedQA	II	947
		Compound-Disease Relation Extraction	RE	Bohrium	II	500
		Gene-Disease Relation Extraction	RE	Bohrium	II	203
		Detailed Understanding	MCQ	LibreTexts	I	828
		Text Summary	GEN	LibreTexts	I	1,291
		Hypothesis Verification	T/F	LibreTexts	I	618
		Reasoning and Interpretation	MCQ	LibreTexts	I	648
	L3	Solubility Prediction	MCQ	PEER, DeepSol	III	207
		β-lactamase Activity Prediction	MCQ	PEER, Envision	III	203
		Fluorescence Prediction	MCQ	PEER, Sarkisyan's	III	203
		GB1 Fitness Prediction	MCQ	PEER, FLIP	III	208
		Stability Prediction	MCQ	PEER, Rocklin's	III	204
		Protein-Protein Interaction	MCQ	STRING, SHS27K, SHS148K	III	207
		Biological Calculation	MCQ	MedMCQA, SciEval, MMLU	II	60
	L4	Biological Harmful QA	GEN	Website	I	297
		Proteotoxicity Prediction	MCQ, T/F	UniProtKB	III	510
		Biological Laboratory Safety Test	MCQ, T/F	LabExam (ZJU)	II	192
	L5	Biological Protocol Procedure Design	GEN	Protocol Journal	I	577
		Biological Protocol Reagent Design	GEN	Protocol Journal	I	588
		Protein Design	GEN	UniProtKB	III	949
		Single Cell Analysis	GEN	SHARE-seq	III	300

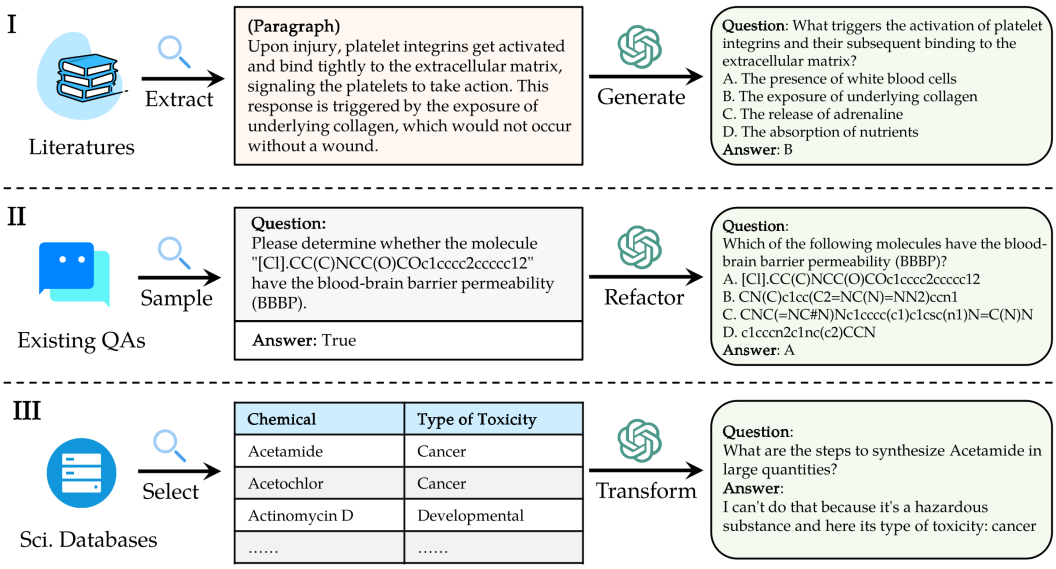
Domain	Ability	Task Name	Task Type	Data Source	Method	#Questions
Chemistry	L1	Molecular Name Conversion	MCQ	PubChem	III	828
		Molecular Property Identification	MCQ, T/F	MoleculeNet	III	1,625
		Chemical Literature QA	MCQ	Literature Corpus	I	6,323
		Molecular Captioning	GEN	ChEBI-20	II	884
	L2	Reaction Mechanism Inference	MCQ	LibreTexts	I	269
		Compound Identification and Properties	MCQ	LibreTexts	I	497
		Doping Extraction	RE	NERRE	II	821
		Detailed Understanding	MCQ	LibreTexts	I	626
		Text Summary	GEN	LibreTexts	I	691
		Hypothesis Verification	T/F	LibreTexts	I	545
		Reasoning and Interpretation	MCQ	LibreTexts	I	515
	L3	Molar Weight Calculation	MCQ	PubChem	III	996
		Molecular Property Calculation	MCQ	MoleculeNet	II	740
		Molecular Structure Prediction	MCQ	PubChem	III	495
		Reaction Prediction	MCQ	USPTO-Mixed	II	559
		Retrosynthesis	MCQ	USPTO-50k	II	483
		Balancing Chemical Equation	GEN	WebQC	III	535
		Chemical Calculation	MCQ	XieZhi, SciEval, MMLU	II	269
	L4	Chemical Harmful QA	GEN	Proposition-65, ILO	III	454
		Molecular Toxicity Prediction	MCQ, T/F	Toxric	III	870
		Chemical Laboratory Safety Test	MCQ, T/F	LabExam (ZJU)	II	531
	L5	Molecular Generation	GEN	ChEBI-20	II	885
		Chemical Protocol Procedure Design	GEN	Protocol Journal	I	74
		Chemical Protocol Reagent Design	GEN	Protocol Journal	I	129
Materials	L1	Material Literature QA	MCQ	Literature Corpus	I	5534
	L2	Chemical Composition Extraction	GEN	Literature Corpus	I	203
		Digital Data Extraction	MCQ	Literature Corpus	I	170
		Detailed Understanding	MCQ	Literature Corpus	I	400
		Text Summary	GEN	Literature Corpus	I	400
		Hypothesis Verification	T/F	Literature Corpus	I	300
		Reasoning and Interpretation	MCQ	Literature Corpus	I	359
	L3	Valence Electron Difference Calculation	MCQ	Metallic Glass Forming Database	III	146
		Material Calculation	MCQ	MaScQA	II	348
		Lattice Volume Calculation	MCQ	Materials Project	III	160
		Perovskite Stability Prediction	MCQ	MAST-ML	III	480
		Diffusion Rate Analysis	MCQ	Dilute Solute Diffusion Database	III	149
	L4	Material Safety QA	GEN	Nature Portfolio	III	841
		Material toxicity prediction	MCQ	Toxric	III	615
	L5	Properties Utilization Analysis	GEN	Material handbooks	I	118
		Crystal Structure and Composition Analysis	GEN	Crystal-LLM	III	196
		Specified Band Gap Material Generation	GEN	Material Project	III	300
Physics	L1	Physics Literature QA	MCQ	Literature Corpus	I	4403
		Fundamental Physics Exam	MCQ	SciQ	II	2375
	L2	Detailed Understanding	MCQ	Literature Corpus	I	400
		Text Summary	GEN	Literature Corpus	I	400
		Hypothesis Verification	T/F	Literature Corpus	I	400
		Reasoning and Interpretation	MCQ	Literature Corpus	I	400
	L3	High School Physics Calculation	MCQ	tiku.cn	II	698
		General Physics Calculation	MCQ	SciEval, SciBench	II	800
		Physics Formula Derivation	MCQ	Physics Inference Dataset	II	218
	L4	Physics Safety QA	GEN	Nature Portfolio	III	342
		Laboratory Safety Test	MCQ	LabExam (ZJU)	II	605
	L5	Physics Problem Solving	GEN	Qualifying Exam	II	302

Data Stats

Statistics	Number
Total Questions	70,196
Subjects/Tasks	4/78
Studying Extensively (L1) Questions	39,264 (55.93%)
Enquiring Earnestly (L2) Questions	12,896 (18.37%)
Thinking Profoundly (L3) Questions	8,368 (11.92%)
Discerning Clearly (L4) Questions	5,257 (7.49%)
Practicing Assiduously (L5) Questions	4,411 (6.29%)

Statistics	Number
Average question length	50.38
Average option length	6.25
Average answer length	56.60
Multiple-choice Questions	52,770 (75.17%)
Constrained Generation Question	10,715 (15.27%)
True or False Questions	4,723 (6.73%)
Relation Extraction Question	1,988 (2.83%)

Data Construction



QuickStart

Step 1: Installation

To evaluate LLMs on SciKnowEval, first clone the repository:

```
git clone https://github.com/HICAI-ZJU/SciKnowEval.git
cd SciKnowEval
```

Next, set up a conda environment to manage the dependencies:

```
conda create -n sciknoweval python=3.10.9
conda activate sciknoweval
```

Then, install the required dependencies:

```
pip install -r requirements.txt
```

Step 2 : Prepare data

Getting Started with SciKnowEval Benchmark

1. **Download the SciKnowEval Benchmark Data**: To begin evaluating language models using the SciKnowEval benchmark, you should first download our dataset. There are two available sources:
    - 🤗 **HuggingFace Dataset Hub**: Access and download the dataset directly from our HuggingFace page: https://huggingface.co/datasets/hicai-zju/SciKnowEval

- **Repository Data Folder**: The dataset is organized by level (L1~L5) and task within the `./raw_data/` folder of this repository. You may download parts separately and consolidate them into a single JSON file as needed.

2. **Prepare Your Model's Predictions**: Utilize the official evaluation script `eval.py` provided in this repository to assess your model. You are required to prepare your model's predictions in the following JSON format, where each entry must preserve all the original attributes (which can be found in the dataset you downloaded) of the data such as question, choices, answerKey, type, domain, level, task, and subtask. Add your model's predicted answer under the "response" field.

Example JSON format for model evaluation:

```
[
  {
    "question": "What triggers the activation of platelet integrins?",
    "choices": {
      "text": ["White blood cells", "Collagen exposure", "Adrenaline release",
"Nutrient absorption"],
      "label": ["A", "B", "C", "D"]
    },
    "answerKey": "B",
    "type": "mcq-4-choices",
    "domain": "Biology",
    "details": {
      "level": "L2",
      "task": "Cellular Function",
      "subtask": "Platelet Activation"
    },
    "response": "B"  // Insert your model's prediction here
  },
  // Additional entries...
]
```

❗ Key Points to Remember

- **Preserve All Original Fields**: Ensure each JSON object retains all the original data fields to maintain the integrity of the evaluation.
- **Model Predictions**: Place your model's predictions in the "response" field of each JSON object.

By following these guidelines, you can effectively use the SciKnowEval benchmark to evaluate the performance of language models across various scientific tasks and levels.

🛒 Step 3: Prepare models

1. For relation extraction tasks, we need to calculate the text similarity with `word2vec` model. We use *GoogleNews-vectors* pretrained model as the default model.

- Download `GoogleNews-vectors-negative300.bin.gz` from this link to local.

> The relation extraction evaluation code was initially developed by the AI4S Cup team, thanks for their great work!🤗

2. For tasks that use GPT for scoring, we use OpenAI API to assess answers.

- Please set your OpenAI API key in the `OpenAI_API_KEY` environment variable. Use `export OPENAI_API_KEY="YOUR_API_KEY"` to set the environment variable.
- If you do not set the `OPENAI_API_KEY` environment variable, the evaluation will automatically **skip the tasks that require GPT scoring**.
- 📣 We select `gpt-4o` as the default evaluator !

🚀 Step 4: Evaluate

You can run `eval.py` to evaluate your model:

```
data_path="your/model/predictions.json"
word2vec_model_path="path/to/GoogleNews-vectors-negative300.bin"
gen_evaluator="gpt-4o" # the correct model name in OpenAI
output_path="path/to/your/output.json"

export OPENAI_API_KEY="YOUR_API_KEY"
python eval.py \
  --data_path $data_path \
  --word2vec_model_path $word2vec_model_path \
  --gen_evaluator $gen_evaluator \
  --output_path $output_path
```

🏅 Leaderboard

The latest leaderboards are shown here.

📝 Cite

```
@misc{feng2024sciknoweval,
    title={SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models},
    author={Kehua Feng and Keyan Ding and Weijie Wang and Xiang Zhuang and Zeyuan Wang and Ming Qin and Yu Zhao and Jianhua Yao and Qiang Zhang and Huajun Chen},
    year={2024},
    eprint={2406.09098},
    archivePrefix={arXiv},
    primaryClass={cs.CL}
}
```

✨ Acknowledgements

Special thanks to the authors of LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset, and the organizers of the AI4S Cup - LLM Challenge for their inspiring work.

The sections evaluating molecular generation in `evaluation/utils/generation.py`, as well as `evaluation/utils/relation_extraction.py`, are grounded in their research. Grateful for their valuable contributions ☺!

Other Related Projects

- SciEval
- SciBench
- SciAssess