Dataset card

1. Our dataset contains 57231 problems in the split of Lean Workbook and 82893 problems in the split of Lean Workbook Plus. We provide the natural language statement, answer, formal statement, and formal proof (if available) for each problem. These data can support autoformalization model training and searching for proofs.
2. We open-source our code at `https://github.com/InternLM/InternLM-Math` and our data at `https://huggingface.co/datasets/InternLM/Lean-Workbook`.
3. Croissant metadata URL: `https://huggingface.co/api/datasets/internlm/Lean-Workbook/croissant`.
4. The license of our dataset is Apache 2.0.
5. We will host our dataset in Huggingface and our code in GitHub. We will maintain this dataset with further improvement.
6. DOI of dataset: 10.57967/hf/2399